Applications of Self-Supervised Learning

2022.03.04

발표자: 이영재

KOREA UNIVERSITY

DMQA

발표자 소개

❖ 이름: 이영재 (Young Jae Lee)

- Data Mining & Quality Analytics Lab
- Ph.D. Candidate (2019.03 ~ Present)
- 지도 교수: 김성범 교수님

❖ 연구 분야

- Deep Reinforcement Learning
- Offline Reinforcement Learning
- Multi-Agent Reinforcement Learning
- Self-Supervised Learning

❖ 연락망

- E-mail: jae601@korea.ac.kr

목차

- ❖ Self-Supervised Learning
- ❖ Applications of Self-Supervised Learning
- ❖ Conclusion

Self-Supervised Learning

❖ Supervised Learning vs. Self-Supervised Learning

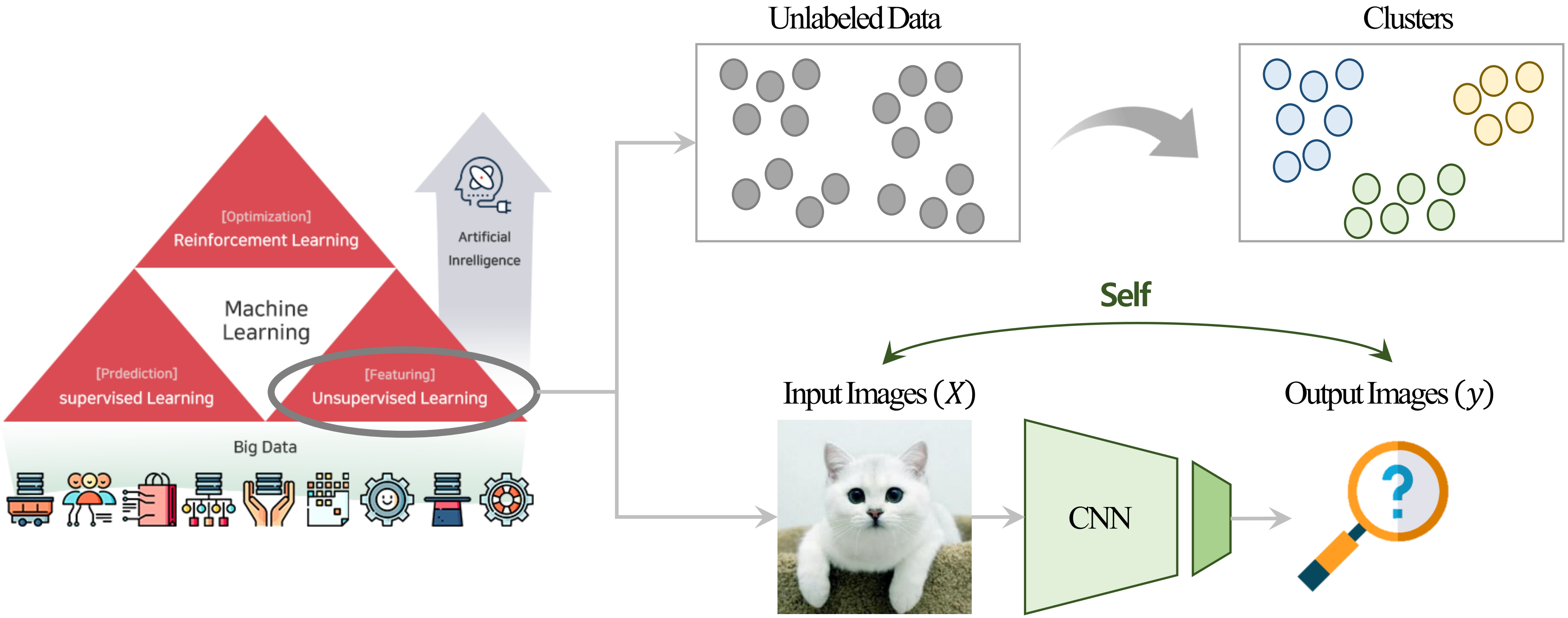
- 공통점: 레이블(y), 사전에 정의한 Task와 같은 것을 예측
- **차이점**: 입력 데이터에 대한 레이블/정답 존재 여부
- **충분한 양의 레이블 정보를 얻는 것은 매우 어려움**

	Supervised Learning	Contrastive Learning
Training Data	(X, Y)	(X)
Model	$\hat{\mathrm{Y}} = F(\mathrm{X})$	$Self\left(\hat{\mathrm{Y}} = F(\mathrm{X})\right)$
Objective	$\left(\mathrm{Y} - \hat{\mathrm{Y}}\right)^2$	$\mathcal{L}_i = -\log \dfrac{exp\left(\frac{i \cdot i^+}{\tau}\right)}{exp\left(\frac{i \cdot i^+}{\tau}\right) + \sum_{k \notin \{i, i^+\}} exp\left(\frac{i \cdot k}{\tau}\right)}$

Self-Supervised Learning

❖ Self-Supervised Learning

- 레이블이 없는(Unlabeled) 데이터를 사용하여 사용자가 **새로운 문제와 정답(Pretext Task)을 정의**하거나 **대조 손실 함수(Contrastive Loss)를 정의**하여 데이터 자체에 대한 이해를 높이는 방법

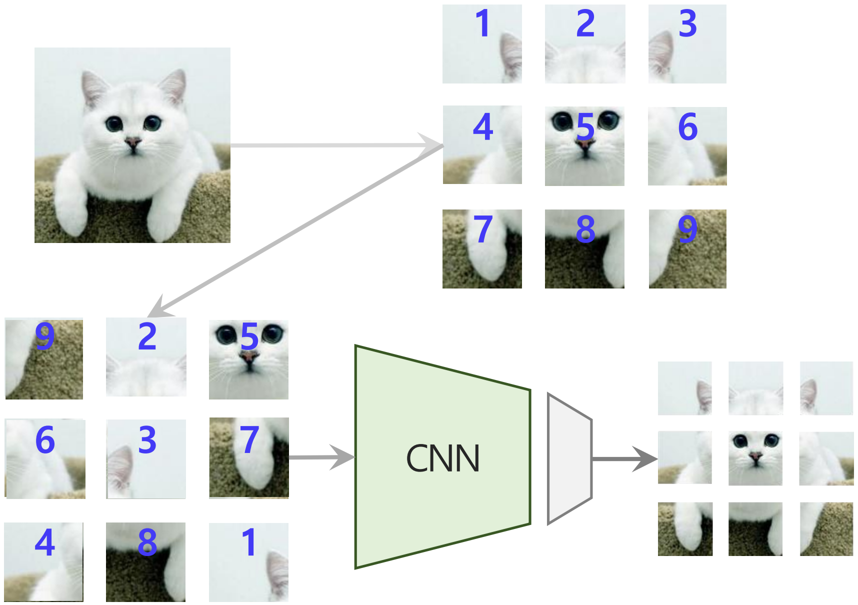


Self-Supervised Learning

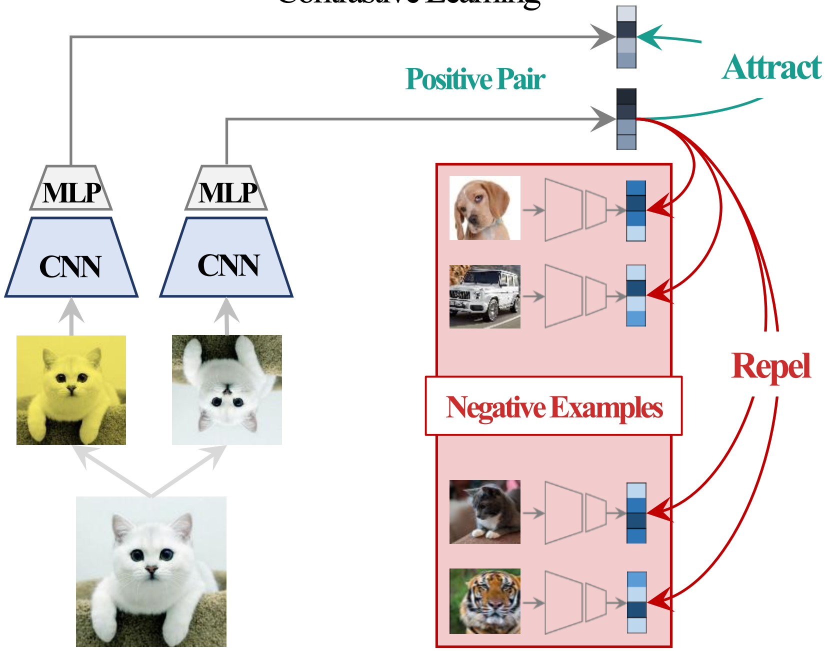
❖ Self-Supervised Learning

- Pretext Tasks: Examplar, Context Prediction, Jigsaw Puzzle, Count, Rotation, …
- Contrastive Learning: MoCo, SimCLR, …
- Non-Contrastive Learning: BYOL

Pretext Task 예: Jigsaw Puzzle



Contrastive Learning



Self-Supervised Learning

❖ MoCo: Momentum Contrast for Unsupervised Visual Representation Learning (CVPR, 2020)

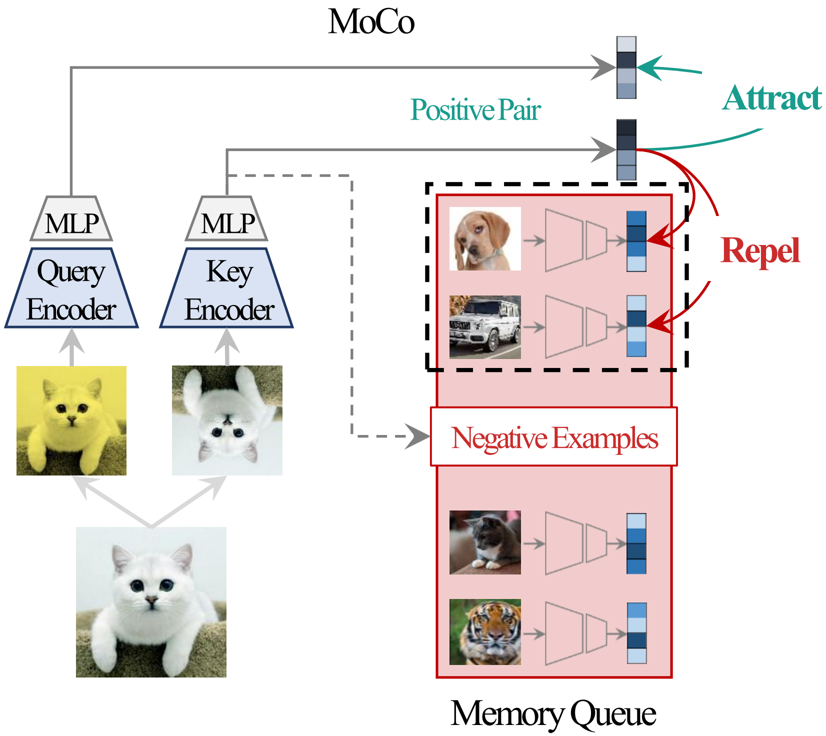
- Memory Queue (First in First Out)
- InfoNCE Loss Function

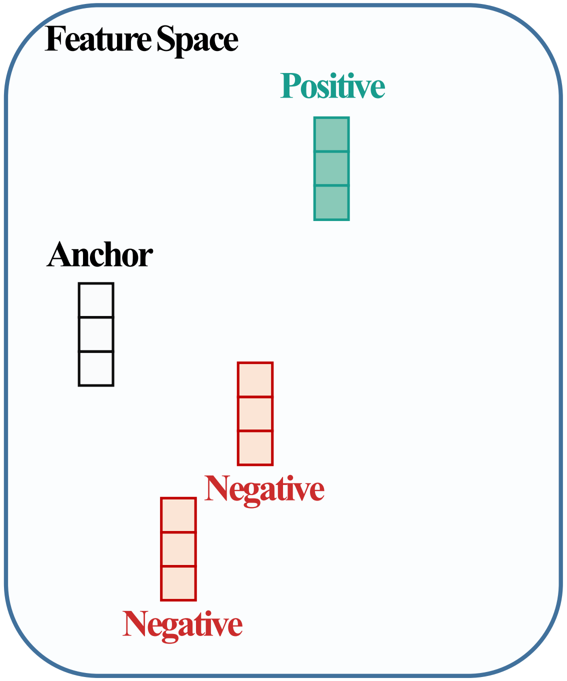
$$\mathcal{L}_i = -log \frac{exp\left(\frac{i \cdot i^+}{\tau}\right)}{exp\left(\frac{i \cdot i^+}{\tau}\right) + \sum_{k \notin \{i, i^+\}} exp\left(\frac{i \cdot k}{\tau}\right)}$$

i: Anchor

i^+: Positive

k: Negative Examples

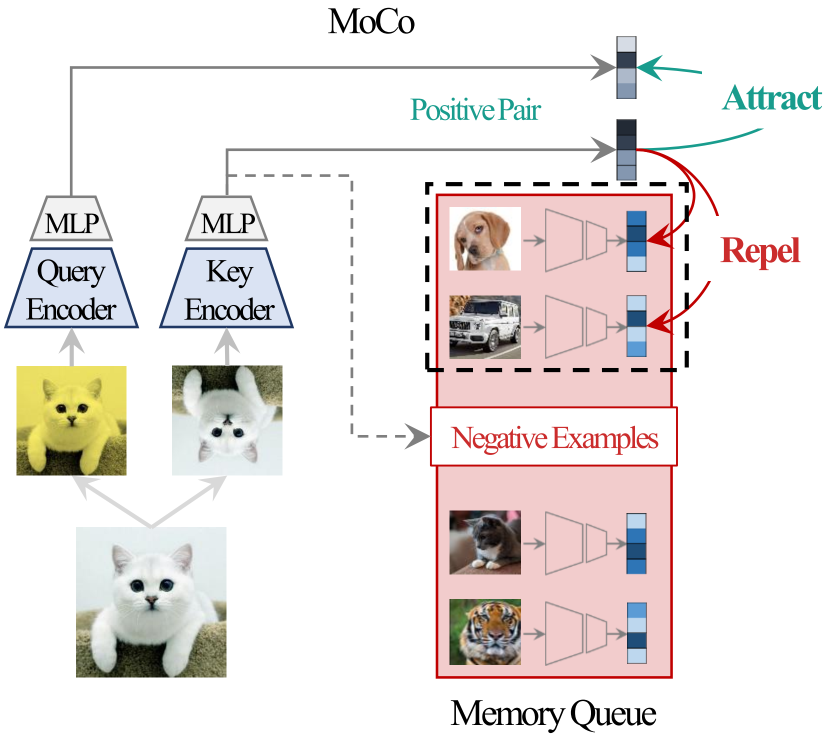




Self-Supervised Learning

❖ MoCo: Momentum Contrast for Unsupervised Visual Representation Learning (CVPR, 2020)

- Momentum Update
- 과거 모델 파라미터들의 가중평균으로 업데이트



Momentum Update
(Key Encoder)

$$\theta_{key} \leftarrow m\theta_{key} + (1-m)\theta_{query}$$

Self-Supervised Learning

❖ SimCLR: A Simple Framework for Contrastive Learning of Visual Representations (ICML, 2020)

- Batch Size (Without Memory Queue)
- Deep Embedding Layer
- InfoNCE Loss Function

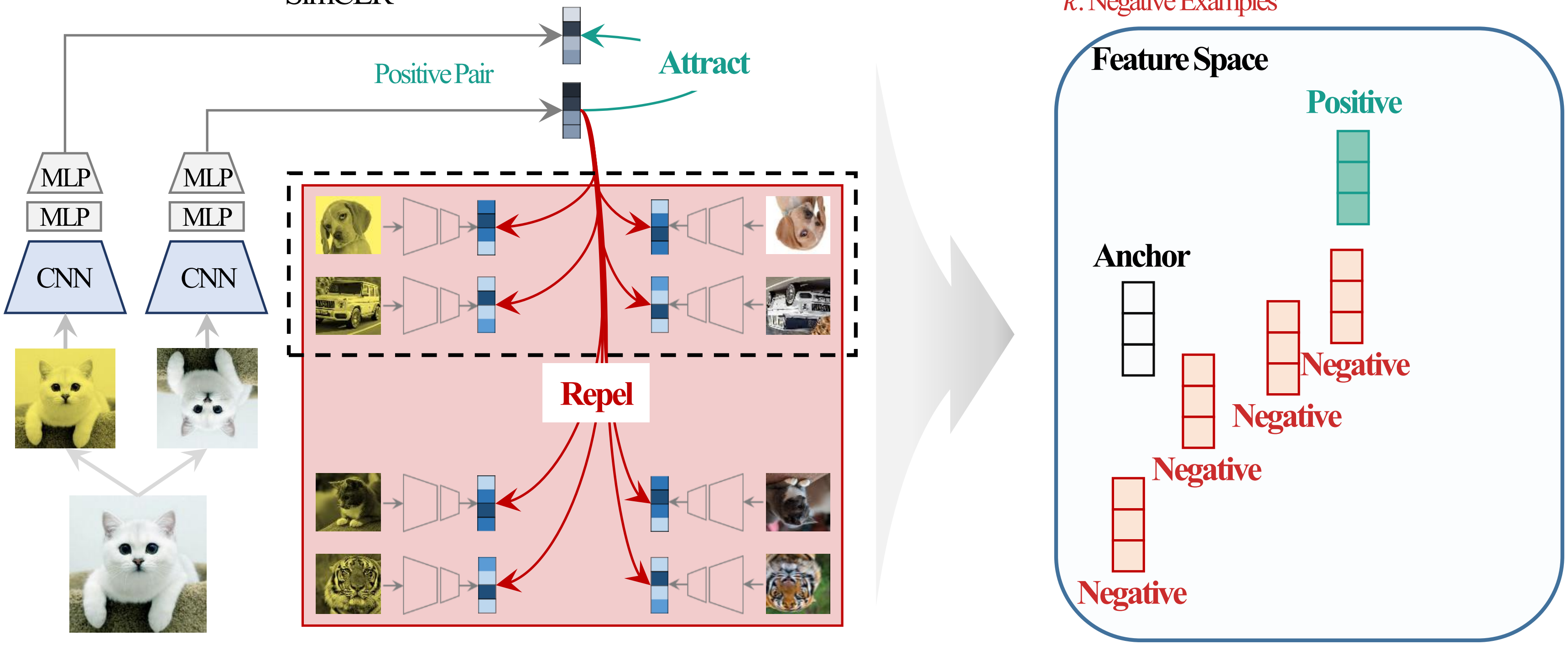
$$\mathcal{L}_i = -log \frac{exp\left(\frac{i \cdot i^+}{\tau}\right)}{exp\left(\frac{i \cdot i^+}{\tau}\right) + \sum_{k \notin \{i, i^+\}} exp\left(\frac{i \cdot k}{\tau}\right)}$$

i: Anchor

i^+: Positive

k: Negative Examples

SimCLR



Negative Examples

Self-Supervised Learning

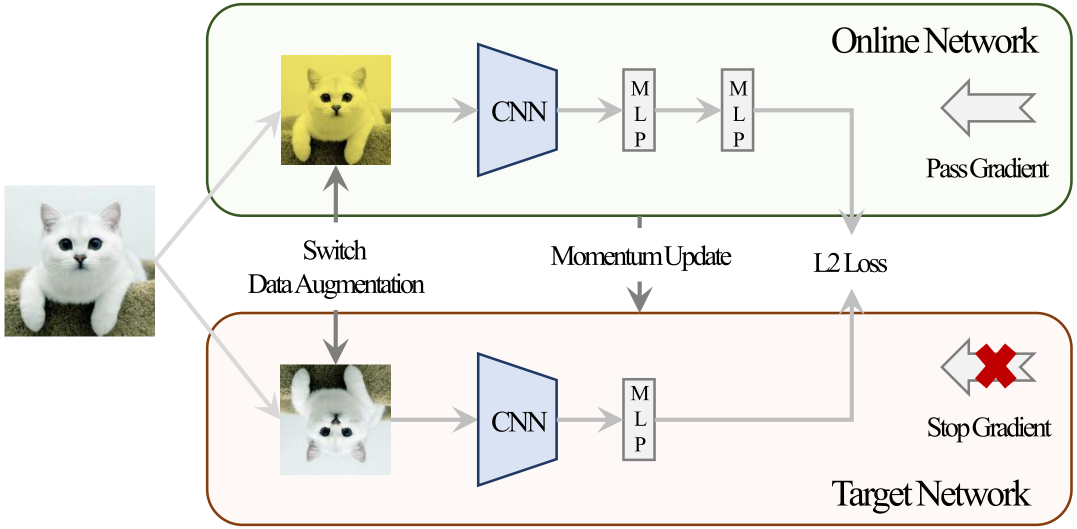
❖ BYOL: Bootstrap Your Own Latent – A New Approach to Self-Supervised Learning (NeurlPS, 2020)

- Positive Pair (Without Negative Examples)
- Momentum Update
- L2 Loss Function

$$\mathcal{L}_{\theta,\delta}^{BYOL} = \left|\left|\overline{q_\theta}(z_\theta) - \bar{z_\delta}'\right|\right|_2^2$$

$$where\ \bar{q_\theta}(z_\theta) \triangleq \frac{q_\theta(z_\theta)}{\left|\left|q_\theta(z_\theta)\right|\right|_2}, \bar{z_\delta}' \triangleq \frac{z_\delta'}{\left|\left|z_\delta'\right|\right|_2}$$

$$Total\ Loss = \mathcal{L}_{\theta,\delta}^{BYOL} + \tilde{\mathcal{L}}_{\theta,\delta}^{BYOL}$$



Momentum Update
(Target Network)

$$\theta_{target} \leftarrow \tau\theta_{target} + (1-\tau)\theta_{online}$$

Applications of Self-Supervised Learning

❖ Multi-Modal Learning

- Contrastive Learning of Medical Visual Representations from Paired Images and Text (arXiv, 2020)

❖ Video

- Spatiotemporal Contrastive Video Representation Learning (CVPR, 2021)

❖ Reinforcement Learning

- Generalization in Reinforcement Learning by Soft Data Augmentation (IEEE ICRA, 2021)

❖ Audio

- Multi-Format Contrastive Learning of Audio Representations (arXiv, 2021)

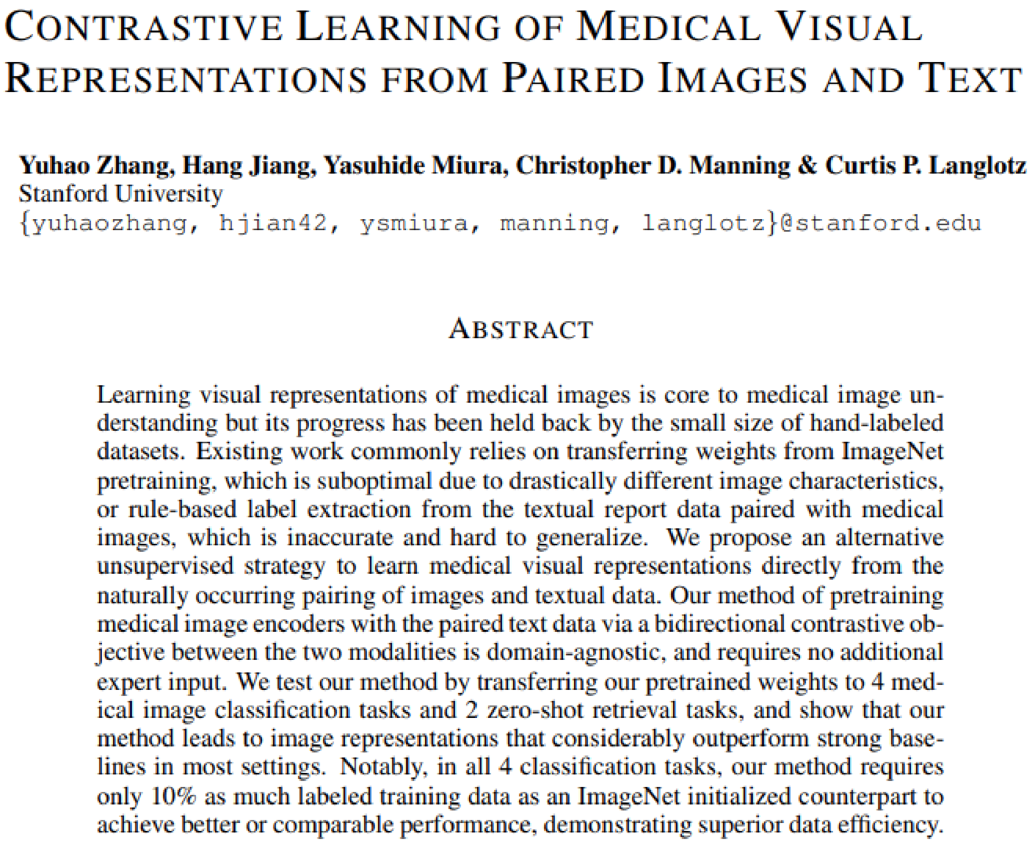
❖ Graph

- Molecular Contrastive Learning of Representations via Graph Neural Networks (arXiv, 2021)

Applications of Self-Supervised Learning

Multi-Modal

- ❖ Contrastive Learning of Medical Visual Representations from Paired Images and Text (arXiv, 2020)
  - Stanford 대학에서 연구하였고 2022년 03월 03일 기준으로 74회 인용



CONTRASTIVE LEARNING OF MEDICAL VISUAL REPRESENTATIONS FROM PAIRED IMAGES AND TEXT

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning & Curtis P. Langlotz
Stanford University
{yuhaozhang, hjian42, ysmiura, manning, langlotz}@stanford.edu

ABSTRACT

Learning visual representations of medical images is core to medical image understanding but its progress has been held back by the small size of hand-labeled datasets. Existing work commonly relies on transferring weights from ImageNet pretraining, which is suboptimal due to drastically different image characteristics, or rule-based label extraction from the textual report data paired with medical images, which is inaccurate and hard to generalize. We propose an alternative unsupervised strategy to learn medical visual representations directly from the naturally occurring pairing of images and textual data. Our method of pretraining medical image encoders with the paired text data via a bidirectional contrastive objective between the two modalities is domain-agnostic, and requires no additional expert input. We test our method by transferring our pretrained weights to 4 medical image classification tasks and 2 zero-shot retrieval tasks, and show that our method leads to image representations that considerably outperform strong baselines in most settings. Notably, in all 4 classification tasks, our method requires only 10% as much labeled training data as an ImageNet initialized counterpart to achieve better or comparable performance, demonstrating superior data efficiency.

Applications of Self-Supervised Learning

Multi-Modal

❖ Contrastive Learning of Medical Visual Representations from Paired Images and Text (arXiv, 2020)

- 의료 이미지의 **시각적 표현을 학습**하는 것은 **의료 이미지 이해**의 핵심
- 기존에는 의료 영상에 대한 레이블을 얻기 위해 두 가지 접근 방식이 존재
  - ✓ 의료 전문가가 직접 만든 레이블
  - ✓ ImageNet 기반의 사전 학습에서 네트워크 가중치를 활용
- 하지만, 의료 영상을 이해하기 위해서는 **매우 세밀한 시각적 특징을 표현**해야 함(Figure 1)
- 본 연구에서는 시각적 표현을 학습하기 위해 **이미지와 텍스트 데이터 쌍**을 사용한 **대조 학습** 방법을 제안

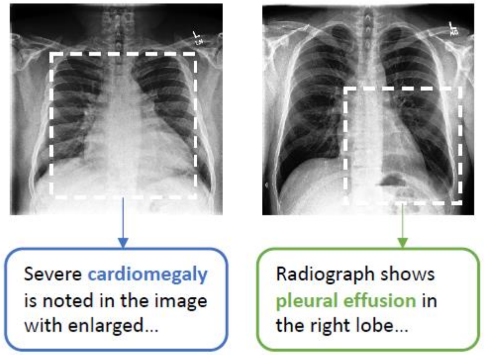

Severe cardiomegaly is noted in the image with enlarged...
Radiograph shows pleural effusion in the right lobe...

Figure 1: Two example chest radiograph images with different abnormality categories, along with sentences from their paired textual report and example views indicative of their characteristics.

Applications of Self-Supervised Learning

Multi-Modal

❖ **Con**trastive **V**isual **R**epresentation Learning from **T**ext (ConVIRT)

- 의료 이미지와 텍스트 데이터 쌍을 사용하여 데이터 자체에 대한 좋은 표현을 학습
  - ✓ (X_v, X_u) : Paired Input
  - ✓ (t_v, t_u) : Data Augmentation (or Transformation) → Crop, Flip, Affine, Color Jitter, Gaussian Blur / Random Sampling
  - ✓ $(\tilde{X}_v, \tilde{X}_u)$: Transformed Input

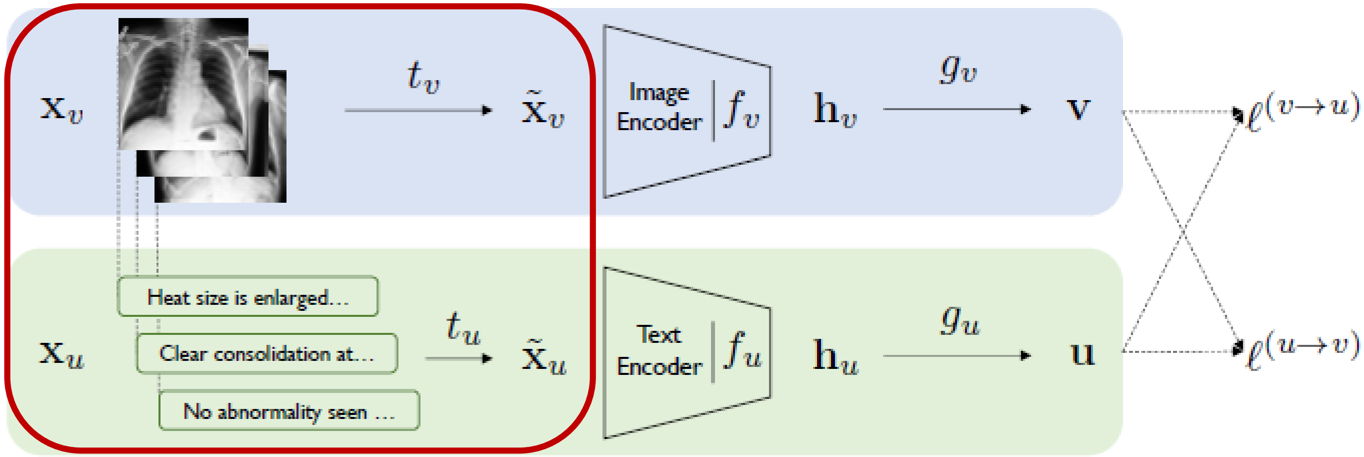


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v \to u)}$ and $\ell^{(u \to v)}$.

Applications of Self-Supervised Learning

Multi-Modal

❖ Contrastive Visual Representation Learning from Text (ConVIRT)

- 의료 이미지와 텍스트 데이터 쌍을 사용하여 데이터 자체에 대한 좋은 표현을 학습
  - ✓ (f_v, f_u): Learning Encoder → f_v는 CNN (ResNet 50), f_u는 BERT 인코더를 사용
  - ✓ (h_v, h_u): Encoder Output
  - ✓ (g_v, g_u): Non-Linear Projection Function
  - ✓ (v, u): Image & Text Representations → $v = g_v\left(f_v(\tilde{X}_v)\right), u = g_u\left(f_u(\tilde{X}_u)\right)$

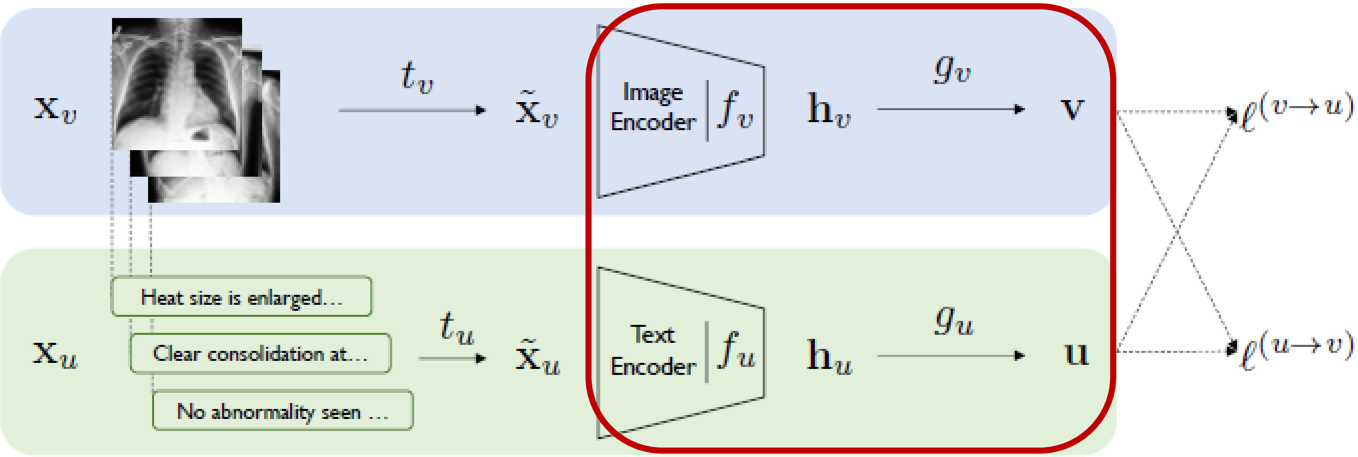


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

Applications of Self-Supervised Learning

Multi-Modal

❖ **Con**trastive **V**isual **R**epresentation Learning from **T**ext (ConVIRT)

- 의료 이미지와 텍스트 데이터 쌍을 사용하여 데이터 자체에 대한 좋은 표현을 학습

✓ Image-to-Text Contrastive Loss: $\ell_i^{(v\rightarrow u)} = -log\frac{\exp(<v_i,u_+>/\tau)}{\exp(<v_i,u_+>/\tau)+\sum_{k\notin\{i,i^+\}}\exp(<v_i,u_k>/\tau)}$

✓ Text-to-Image Contrastive Loss: $\ell_i^{(u\rightarrow v)} = -log\frac{\exp(<u_i,v_+>/\tau)}{\exp(<u_i,v_+>/\tau)+\sum_{k\notin\{i,i^+\}}\exp(<u_i,v_k>/\tau)}$

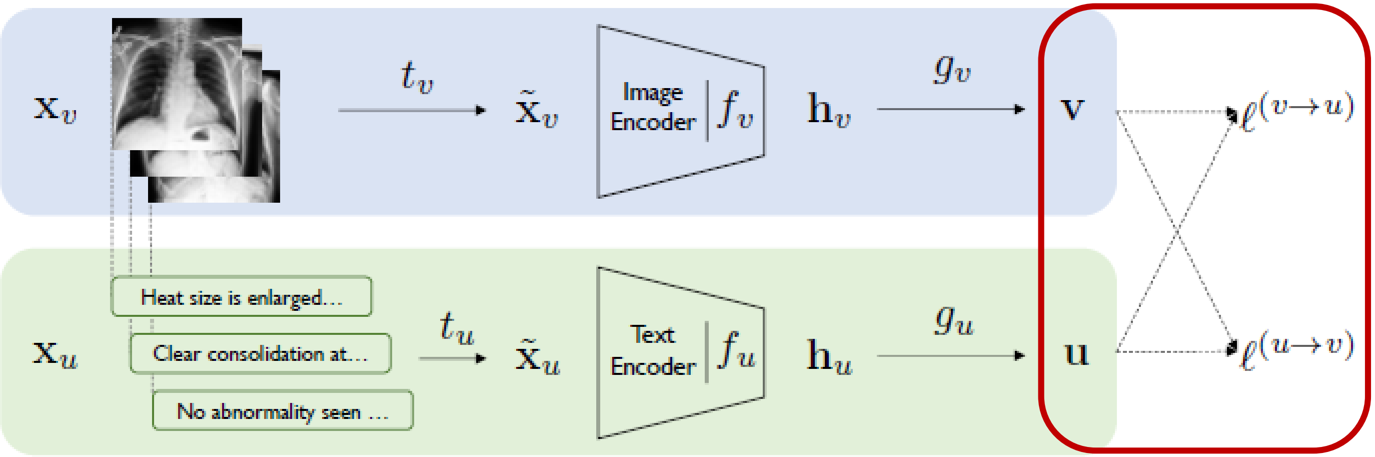


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\rightarrow u)}$ and $\ell^{(u\rightarrow v)}$.

Applications of Self-Supervised Learning

Multi-Modal

❖ **Con**trastive **V**isual **R**epresentation Learning from **T**ext (ConVIRT)

- 의료 이미지와 텍스트 데이터 쌍을 사용하여 데이터 자체에 대한 좋은 표현을 학습
  - ✓ Total Training Loss: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda\ell_i^{(v\to u)} + (1-\lambda)\ell_i^{(u\to v)}\right), \lambda \in [0,1]$

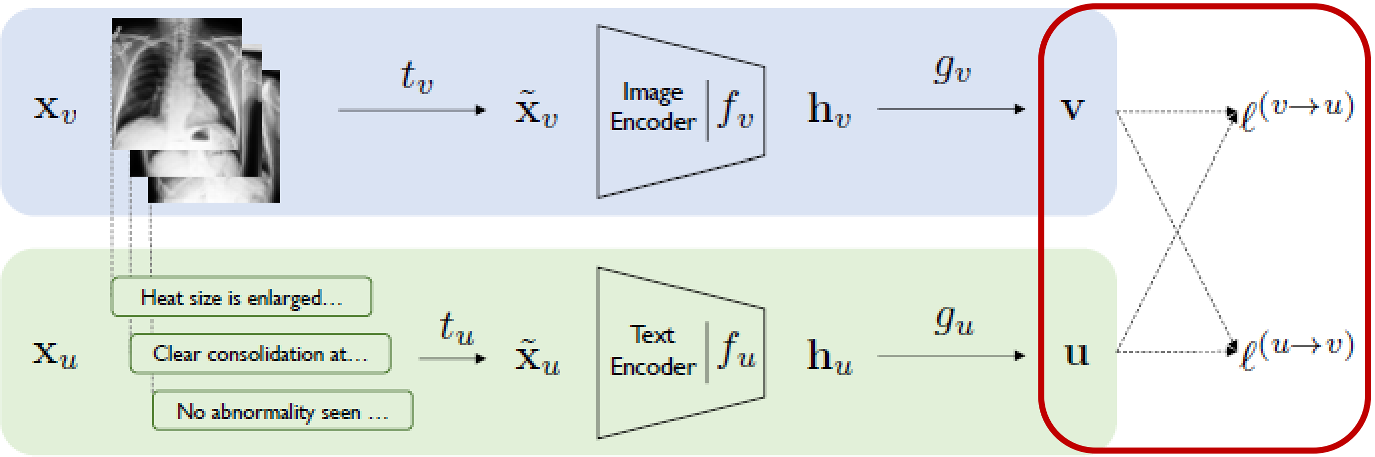


Figure 2: Overview of our ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

Applications of Self-Supervised Learning

Multi-Modal

❖ Experiments – Image Classification

- Four Representative Medical Image Classification Tasks
  - ✓ RSNA Pneumonia Detection: Binary Classification (*Pneumonia or a normal*)
  - ✓ CheXpert: Multi-Label Binary Classification (*Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion*)
  - ✓ COVIDx: Multi-Class Classification (*COVID19, non-COVID Pneumonia, Normal Categories*)
  - ✓ MURA: Binary Classification (*Abnormal, Normal*)
- COVIDx에 대한 Test Accuracy / 이전 방법론들과 동일하게 AUC Metric으로 비교
- Data Efficiency를 비교하기 위해 1%, 10%, All 훈련 데이터로 학습한 이미지 인코더를 평가

Applications of Self-Supervised Learning

Multi-Modal

- ❖ Experiments – Image Classification
  - Four Representative Medical Image Classification Tasks

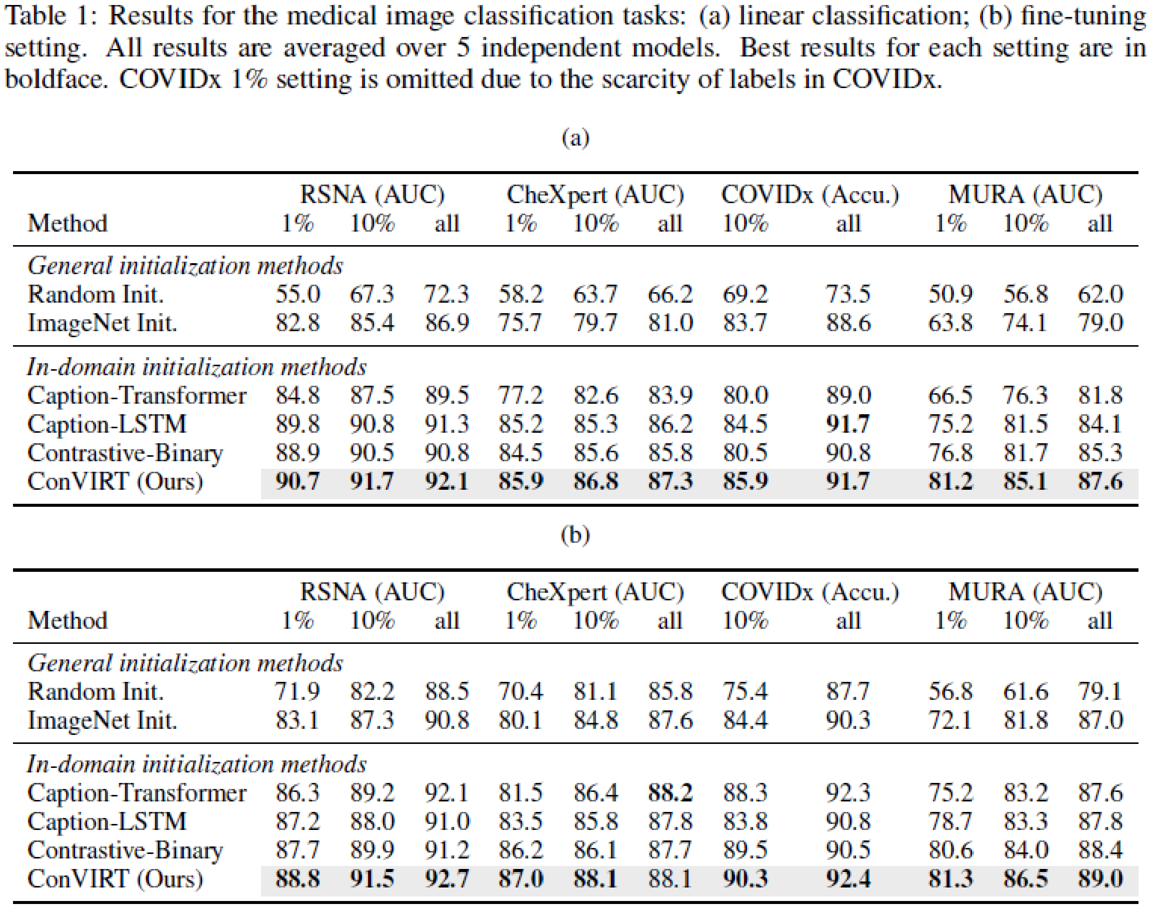


Table 1: Results for the medical image classification tasks: (a) linear classification; (b) fine-tuning setting. All results are averaged over 5 independent models. Best results for each setting are in boldface. COVIDx 1% setting is omitted due to the scarcity of labels in COVIDx.

(a)

	RSNA (AUC)			CheXpert (AUC)			COVIDx (Accu.)		MURA (AUC)		
Method	1%	10%	all	1%	10%	all	10%	all	1%	10%	all
General initialization methods											
Random Init.	55.0	67.3	72.3	58.2	63.7	66.2	69.2	73.5	50.9	56.8	62.0
ImageNet Init.	82.8	85.4	86.9	75.7	79.7	81.0	83.7	88.6	63.8	74.1	79.0
In-domain initialization methods											
Caption-Transformer	84.8	87.5	89.5	77.2	82.6	83.9	80.0	89.0	66.5	76.3	81.8
Caption-LSTM	89.8	90.8	91.3	85.2	85.3	86.2	84.5	**91.7**	75.2	81.5	84.1
Contrastive-Binary	88.9	90.5	90.8	84.5	85.6	85.8	80.5	90.8	76.8	81.7	85.3
ConVIRT (Ours)	**90.7**	**91.7**	**92.1**	**85.9**	**86.8**	**87.3**	**85.9**	**91.7**	**81.2**	**85.1**	**87.6**

(b)

	RSNA (AUC)			CheXpert (AUC)			COVIDx (Accu.)		MURA (AUC)		
Method	1%	10%	all	1%	10%	all	10%	all	1%	10%	all
General initialization methods											
Random Init.	71.9	82.2	88.5	70.4	81.1	85.8	75.4	87.7	56.8	61.6	79.1
ImageNet Init.	83.1	87.3	90.8	80.1	84.8	87.6	84.4	90.3	72.1	81.8	87.0
In-domain initialization methods											
Caption-Transformer	86.3	89.2	92.1	81.5	86.4	**88.2**	88.3	92.3	75.2	83.2	87.6
Caption-LSTM	87.2	88.0	91.0	83.5	85.8	87.8	83.8	90.8	78.7	83.3	87.8
Contrastive-Binary	87.7	89.9	91.2	86.2	86.1	87.7	89.5	90.5	80.6	84.0	88.4
ConVIRT (Ours)	**88.8**	**91.5**	**92.7**	**87.0**	**88.1**	88.1	**90.3**	**92.4**	**81.3**	**86.5**	**89.0**

Applications of Self-Supervised Learning

Multi-Modal

❖ Experiments – Comparisons to Image-only Contrastive Learning

- SimCLR, MoCo v2: 대표적인 이미지 기반의 대조 학습 방법
- 기존 방법들은 이미지의 좋은 표현을 학습하는데 효과적이지만, 의료 영상을 이해하기 위해서는 더욱 세밀한 시각적 특징을 표현해야 함
- 제안하는 ConVIRT는 이미지와 텍스트 쌍을 동시에 사용함으로써 의료 이미지 이해도를 높임

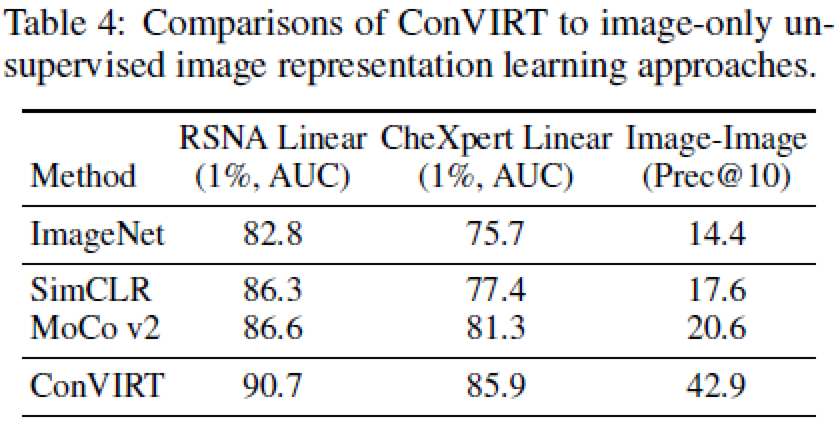


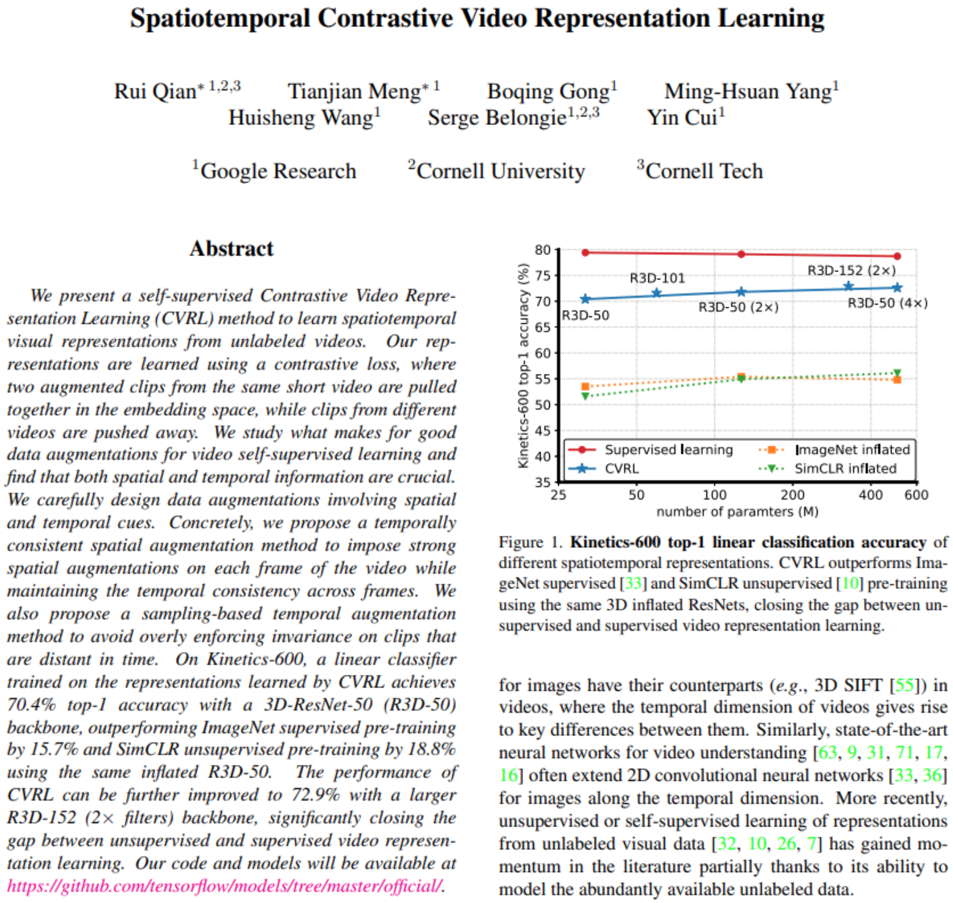
Table 4: Comparisons of ConVIRT to image-only unsupervised image representation learning approaches.

Method	RSNA Linear (1%, AUC)	CheXpert Linear (1%, AUC)	Image-Image (Prec@10)
ImageNet	82.8	75.7	14.4
SimCLR	86.3	77.4	17.6
MoCo v2	86.6	81.3	20.6
ConVIRT	90.7	85.9	42.9

Applications of Self-Supervised Learning

Video

- ❖ Spatiotemporal Contrastive Video Representation Learning (CVPR, 2021)
    - Google Research, Cornell 대학에서 연구하였고 2022년 03월 03일 기준 약 114회 인용



Spatiotemporal Contrastive Video Representation Learning

Rui Qian* [1,2,3] Tianjian Meng* [1] Boqing Gong[1] Ming-Hsuan Yang[1]
Huisheng Wang[1] Serge Belongie[1,2,3] Yin Cui[1]

[1]Google Research [2]Cornell University [3]Cornell Tech

Abstract

We present a self-supervised Contrastive Video Representation Learning (CVRL) method to learn spatiotemporal visual representations from unlabeled videos. Our representations are learned using a contrastive loss, where two augmented clips from the same short video are pulled together in the embedding space, while clips from different videos are pushed away. We study what makes for good data augmentations for video self-supervised learning and find that both spatial and temporal information are crucial. We carefully design data augmentations involving spatial and temporal cues. Concretely, we propose a temporally consistent spatial augmentation method to impose strong spatial augmentations on each frame of the video while maintaining the temporal consistency across frames. We also propose a sampling-based temporal augmentation method to avoid overly enforcing invariance on clips that are distant in time. On Kinetics-600, a linear classifier trained on the representations learned by CVRL achieves 70.4% top-1 accuracy with a 3D-ResNet-50 (R3D-50) backbone, outperforming ImageNet supervised pre-training by 15.7% and SimCLR unsupervised pre-training by 18.8% using the same inflated R3D-50. The performance of CVRL can be further improved to 72.9% with a larger R3D-152 (2× filters) backbone, significantly closing the gap between unsupervised and supervised video representation learning. Our code and models will be available at https://github.com/tensorflow/models/tree/master/official/.

Figure 1. **Kinetics-600 top-1 linear classification accuracy** of different spatiotemporal representations. CVRL outperforms ImageNet supervised [33] and SimCLR unsupervised [10] pre-training using the same 3D inflated ResNets, closing the gap between unsupervised and supervised video representation learning.

for images have their counterparts (*e.g.*, 3D SIFT [55]) in videos, where the temporal dimension of videos gives rise to key differences between them. Similarly, state-of-the-art neural networks for video understanding [63, 9, 31, 71, 17, 16] often extend 2D convolutional neural networks [33, 36] for images along the temporal dimension. More recently, unsupervised or self-supervised learning of representations from unlabeled visual data [32, 10, 26, 7] has gained momentum in the literature partially thanks to its ability to model the abundantly available unlabeled data.

Applications of Self-Supervised Learning

Video

❖ Spatiotemporal Contrastive Video Representation Learning (CVPR, 2021)

- 영상의 **시공간적 표현을 학습**하는 것이 **영상 이해**의 핵심
- 이미지의 표현을 학습하는 **SimCLR 방법 적용**
- 단, 영상 내 시간적으로 일관된 데이터 증강 기법을 적용해야 함(**Figure 3**)
- 본 연구에서는 영상의 시공간적 표현을 학습하기 위해 **시간적으로 일관된 증강 기법**을 적용한 **영상 대조 학습** 방법을 제안

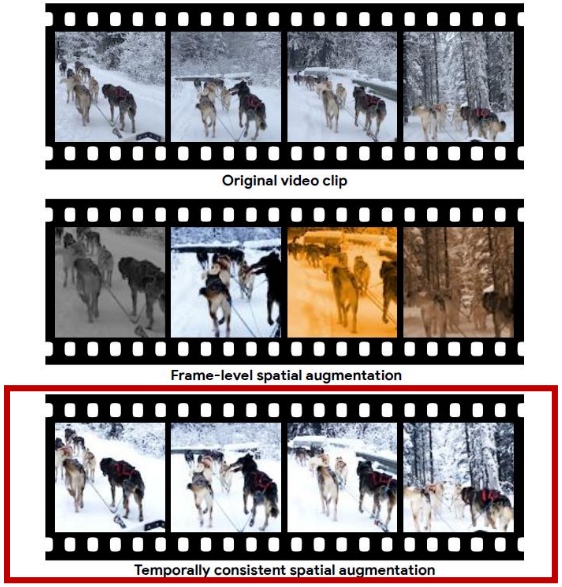

Original video clip
Frame-level spatial augmentation
Temporally consistent spatial augmentation

Figure 3. **Illustration of temporally consistent spatial augmentation.** The middle row indicates frame-level spatial augmentations without temporal consistency which would be detrimental to the video representation learning.

Applications of Self-Supervised Learning

Video

- ❖ Self-Supervised **C**ontrastive **V**ideo **R**epresentation **L**earning (CVRL)
  - 영상 표현 학습을 위해 SimCLR 적용 – Positive Pair와 Negative Examples 정의
    - ✓ 서로 다른 영상(Video 1, Video 2) 선택
    - ✓ 한 영상 내 선택된 서로 다른 시점의 두 Clip은 Positive Pair, 다른 영상 내 선택된 두 Clip은 Negative Examples

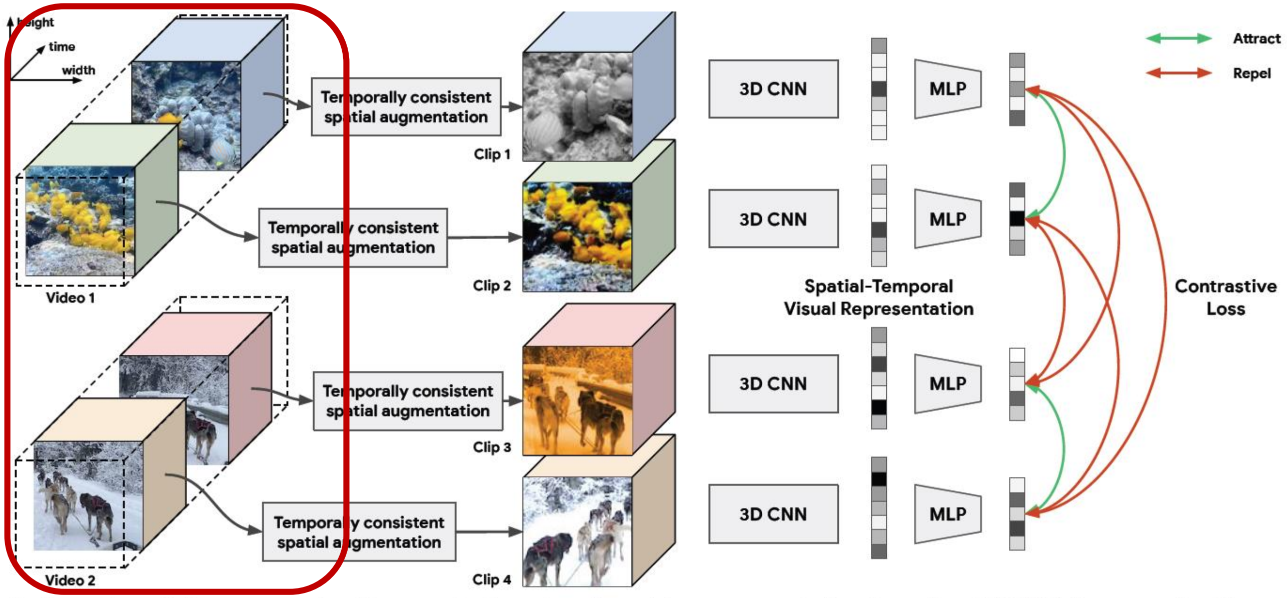
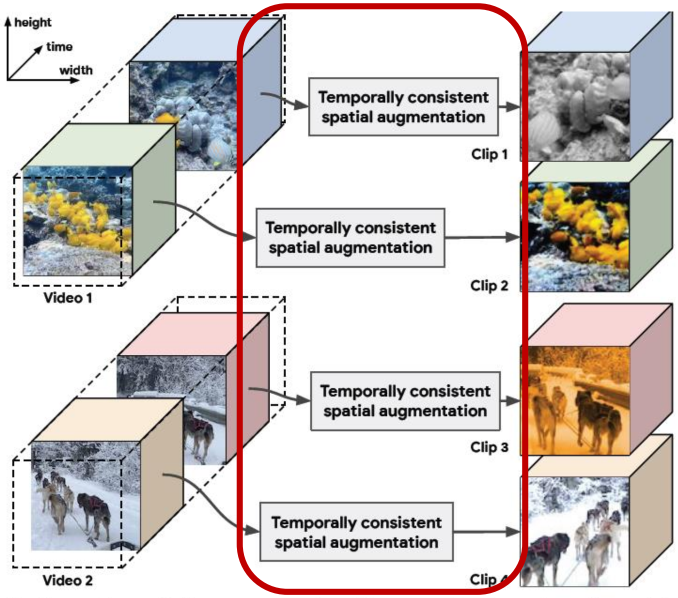

height
time
width
Video 1
Video 2
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Clip 1
Clip 2
Clip 3
Clip 4
3D CNN
3D CNN
3D CNN
3D CNN
MLP
MLP
MLP
MLP
Spatial-Temporal Visual Representation
Attract
Repel
Contrastive Loss

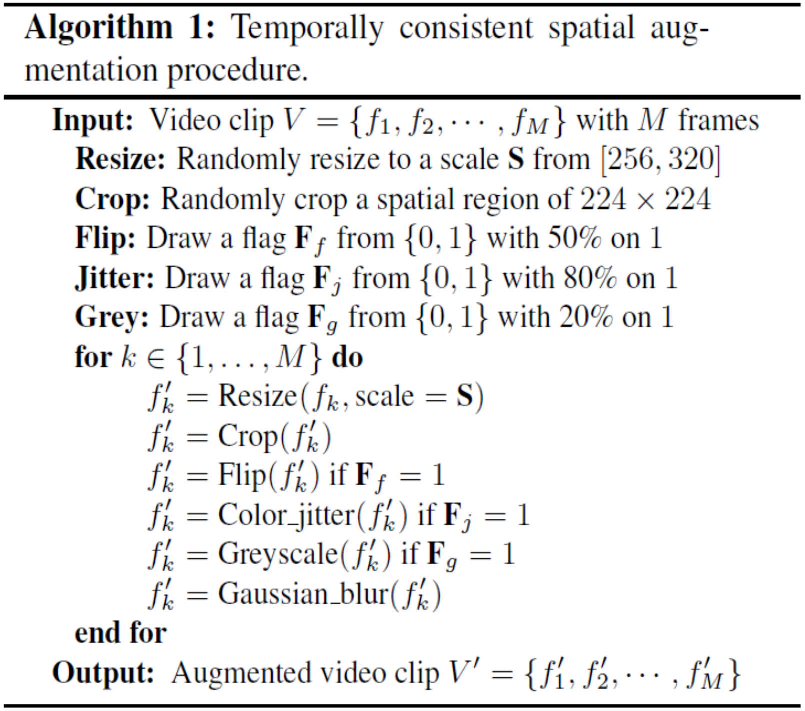
Figure 2. **Overview of the proposed self-supervised contrastive video representation learning (CVRL) framework.** From a short video, we randomly sample 2 clips with the same length. We then apply a temporally consistent spatial augmentation to each of the video clips and feed it to a 3D backbone with an MLP head. The contrastive loss is used to train the network to attract clips from the same video and repel clips from different videos in the embedding space.

Applications of Self-Supervised Learning

Video

- Self-Supervised **C**ontrastive **V**ideo **R**epresentation **L**earning (CVRL)
  - 영상 표현 학습을 위해 SimCLR 적용 – 데이터 증강 기법
    - ✓ Positive Pair, Negative Examples에 시간적으로 일관된 데이터 증강 기법 적용


height
time
width
Temporally consistent spatial augmentation
Clip 1
Temporally consistent spatial augmentation
Clip 2
Video 1
Temporally consistent spatial augmentation
Clip 3
Temporally consistent spatial augmentation
Clip 4
Video 2



Algorithm 1: Temporally consistent spatial augmentation procedure.

```
Input: Video clip $V = \{f_1, f_2, \cdots, f_M\}$ with $M$ frames
  Resize: Randomly resize to a scale $\mathbf{S}$ from $[256, 320]$
  Crop: Randomly crop a spatial region of $224 \times 224$
  Flip: Draw a flag $\mathbf{F}_f$ from $\{0, 1\}$ with 50% on 1
  Jitter: Draw a flag $\mathbf{F}_j$ from $\{0, 1\}$ with 80% on 1
  Grey: Draw a flag $\mathbf{F}_g$ from $\{0, 1\}$ with 20% on 1
  for $k \in \{1, \ldots, M\}$ do
      $f'_k = \text{Resize}(f_k, \text{scale} = \mathbf{S})$
      $f'_k = \text{Crop}(f'_k)$
      $f'_k = \text{Flip}(f'_k)$ if $\mathbf{F}_f = 1$
      $f'_k = \text{Color\_jitter}(f'_k)$ if $\mathbf{F}_j = 1$
      $f'_k = \text{Greyscale}(f'_k)$ if $\mathbf{F}_g = 1$
      $f'_k = \text{Gaussian\_blur}(f'_k)$
  end for
Output: Augmented video clip $V' = \{f'_1, f'_2, \cdots, f'_M\}$
```

Figure 2. **Overview of the proposed self-supervised contrastive video representation learning (CVRL) framework.** From a short video, we randomly sample 2 clips with the same length. We then apply a temporally consistent spatial augmentation to each of the video clips and feed it to a 3D backbone with an MLP head. The contrastive loss is used to train the network to attract clips from the same video and repel clips from different videos in the embedding space.

Applications of Self-Supervised Learning

Video

- Self-Supervised **C**ontrastive **V**ideo **R**epresentation **L**earning (CVRL)
  - 영상 표현 학습을 위해 SimCLR 적용 – 영상 요약
    - ✓ 시공간적 영상을 효과적으로 처리할 수 있는 3D CNN 적용
    - ✓ 3D CNN으로 요약된 특징을 MLP로 한번 더 요약

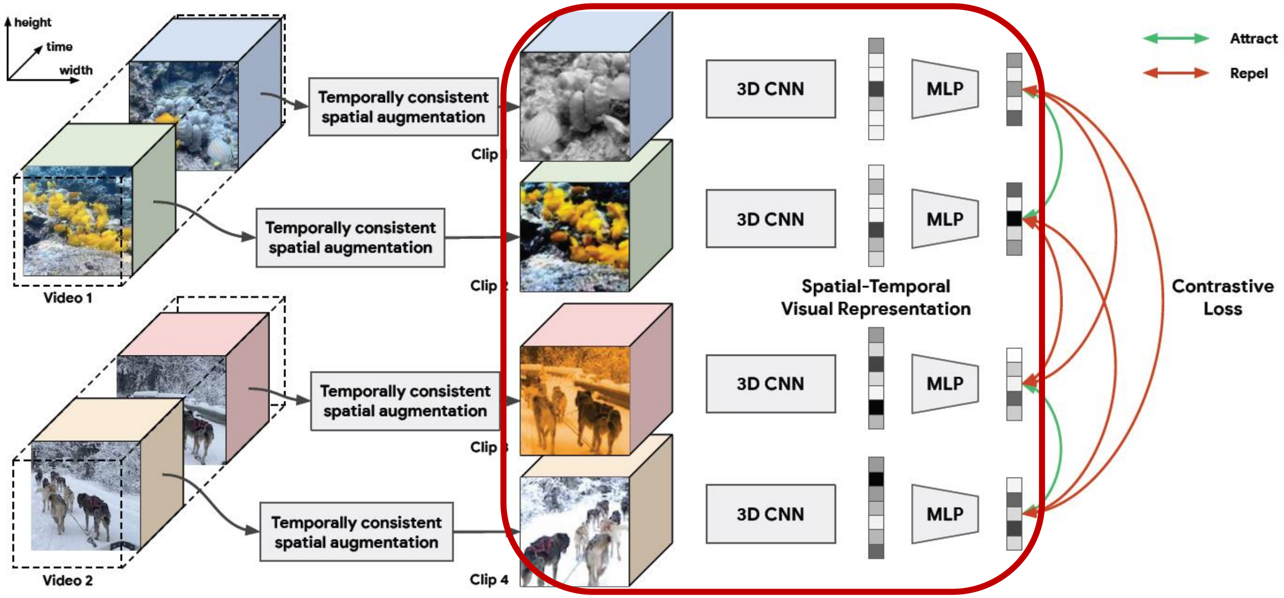

height
time
width
Video 1
Video 2
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Clip 1
Clip 2
Clip 3
Clip 4
3D CNN
3D CNN
3D CNN
3D CNN
MLP
MLP
MLP
MLP
Spatial-Temporal Visual Representation
Attract
Repel
Contrastive Loss

Figure 2. **Overview of the proposed self-supervised contrastive video representation learning (CVRL) framework.** From a short video, we randomly sample 2 clips with the same length. We then apply a temporally consistent spatial augmentation to each of the video clips and feed it to a 3D backbone with an MLP head. The contrastive loss is used to train the network to attract clips from the same video and repel clips from different videos in the embedding space.

Applications of Self-Supervised Learning

Video

- ❖ Self-Supervised **C**ontrastive **V**ideo **R**epresentation **L**earning (CVRL)
  - 영상 표현 학습을 위해 SimCLR 적용 – InfoNCE Loss Function
    - ✓ MLP로 요약된 벡터들을 사용하여 대조 학습 수행

$$\mathcal{L}_i = -\log \frac{exp\left(\frac{i \cdot i^{+}}{\tau}\right)}{exp\left(\frac{i \cdot i^{+}}{\tau}\right) + \sum_{k \notin \{i, i^{+}\}} exp\left(\frac{i \cdot k}{\tau}\right)}$$

i: Anchor

i^{+}: Positive

k: Negative Examples

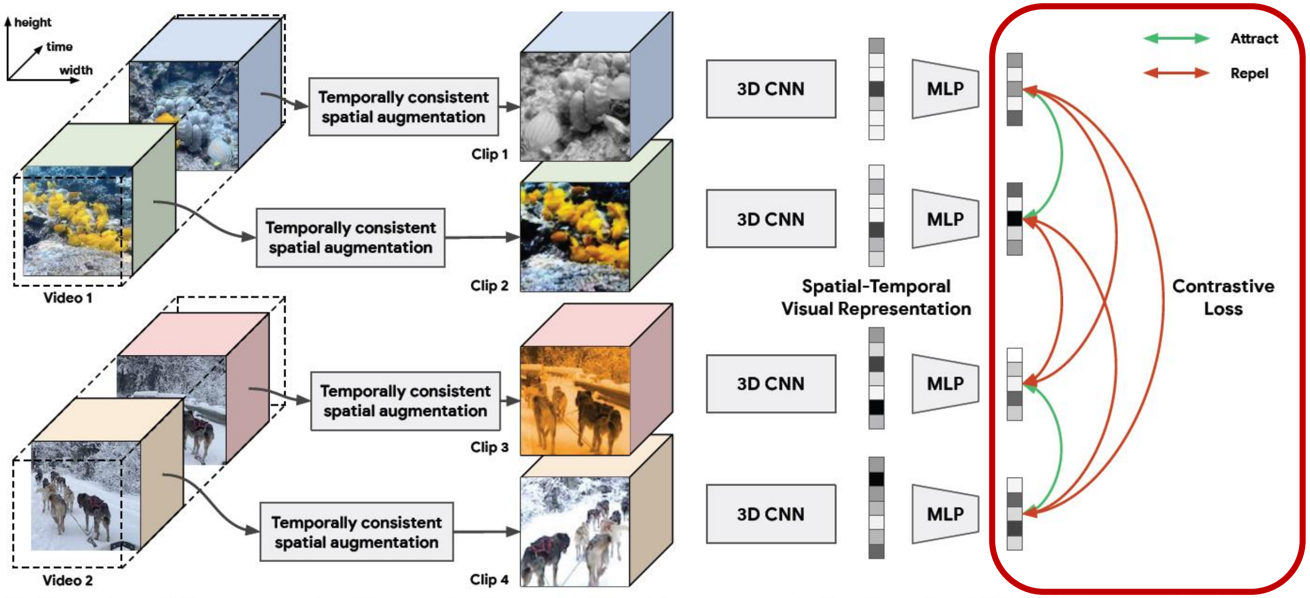

height
time
width
Video 1
Video 2
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Temporally consistent spatial augmentation
Clip 1
Clip 2
Clip 3
Clip 4
3D CNN
3D CNN
3D CNN
3D CNN
MLP
MLP
MLP
MLP
Spatial-Temporal Visual Representation
Attract
Repel
Contrastive Loss

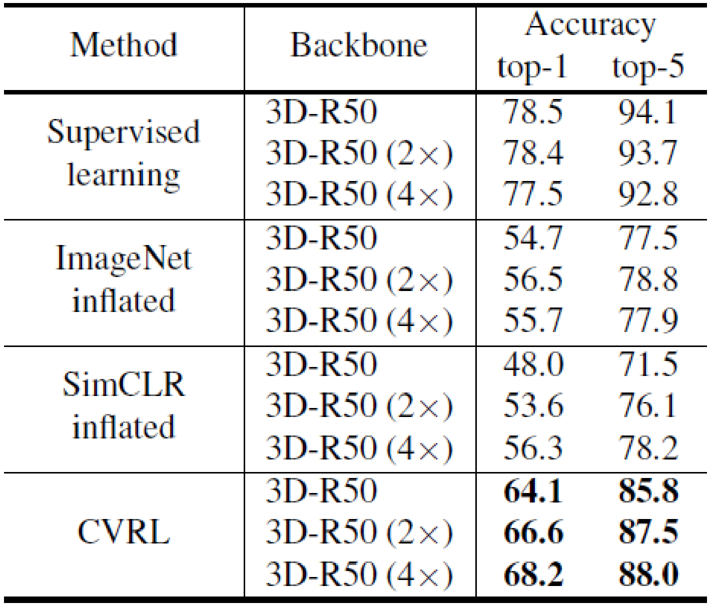
Figure 2. **Overview of the proposed self-supervised contrastive video representation learning (CVRL) framework.** From a short video, we randomly sample 2 clips with the same length. We then apply a temporally consistent spatial augmentation to each of the video clips and feed it to a 3D backbone with an MLP head. The contrastive loss is used to train the network to attract clips from the same video and repel clips from different videos in the embedding space.

Applications of Self-Supervised Learning

Video

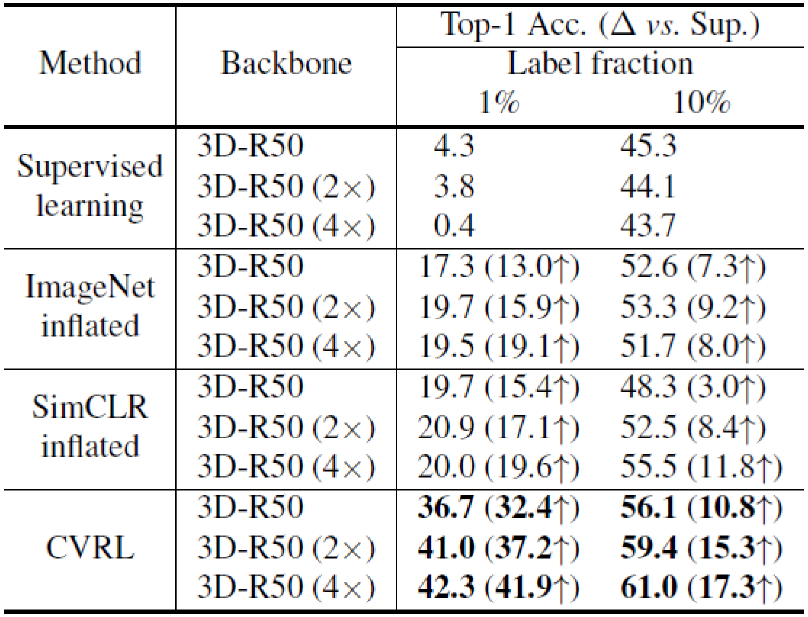
❖ Experiments – Video Classification

- Kinetics-600: Multi-Class Classification (600가지 액션 클래스)
- Supervised Learning / Semi-Supervised Learning에 대한 결과



Method	Backbone	Accuracy top-1	Accuracy top-5
Supervised learning	3D-R50	78.5	94.1
	3D-R50 (2×)	78.4	93.7
	3D-R50 (4×)	77.5	92.8
ImageNet inflated	3D-R50	54.7	77.5
	3D-R50 (2×)	56.5	78.8
	3D-R50 (4×)	55.7	77.9
SimCLR inflated	3D-R50	48.0	71.5
	3D-R50 (2×)	53.6	76.1
	3D-R50 (4×)	56.3	78.2
CVRL	3D-R50	**64.1**	**85.8**
	3D-R50 (2×)	**66.6**	**87.5**
	3D-R50 (4×)	**68.2**	**88.0**

Table 3. **Main results on Kinetics-600.** CVRL shows its effectiveness by surpassing both ImageNet pre-training inflated weights and SimCLR inflated weights by large margins on various network architectures.



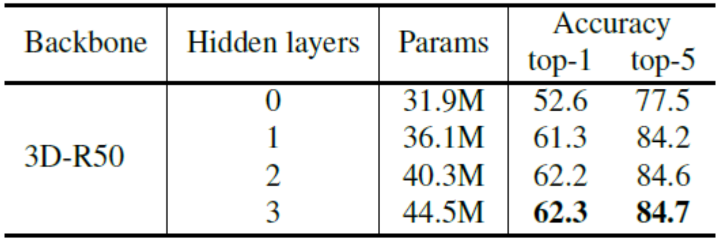
Method	Backbone	Top-1 Acc. (Δ *vs.* Sup.) Label fraction 1%	10%
Supervised learning	3D-R50	4.3	45.3
	3D-R50 (2×)	3.8	44.1
	3D-R50 (4×)	0.4	43.7
ImageNet inflated	3D-R50	17.3 (13.0↑)	52.6 (7.3↑)
	3D-R50 (2×)	19.7 (15.9↑)	53.3 (9.2↑)
	3D-R50 (4×)	19.5 (19.1↑)	51.7 (8.0↑)
SimCLR inflated	3D-R50	19.7 (15.4↑)	48.3 (3.0↑)
	3D-R50 (2×)	20.9 (17.1↑)	52.5 (8.4↑)
	3D-R50 (4×)	20.0 (19.6↑)	55.5 (11.8↑)
CVRL	3D-R50	**36.7 (32.4↑)**	**56.1 (10.8↑)**
	3D-R50 (2×)	**41.0 (37.2↑)**	**59.4 (15.3↑)**
	3D-R50 (4×)	**42.3 (41.9↑)**	**61.0 (17.3↑)**

Table 4. **Semi-supervised learning results on Kinetics-600.** When fine-tuning the entire network with only 1% and 10% labeled data, CVRL outperforms supervised learning, ImageNet pre-training and SimCLR pre-training significantly.

Applications of Self-Supervised Learning

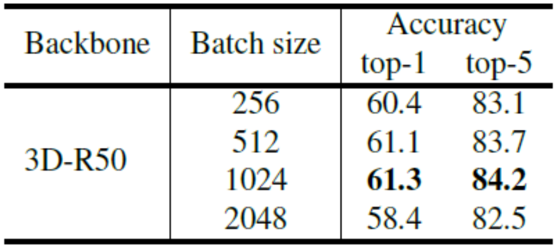
Video

- ❖ Experiments – Ablation Studies
  - 기존 SimCLR는 MLP Layer 수, Batch Size, Epoch에 따라 성능 차이가 존재
  - 제안하는 CVRL로 최적의 하이퍼파라미터 선택



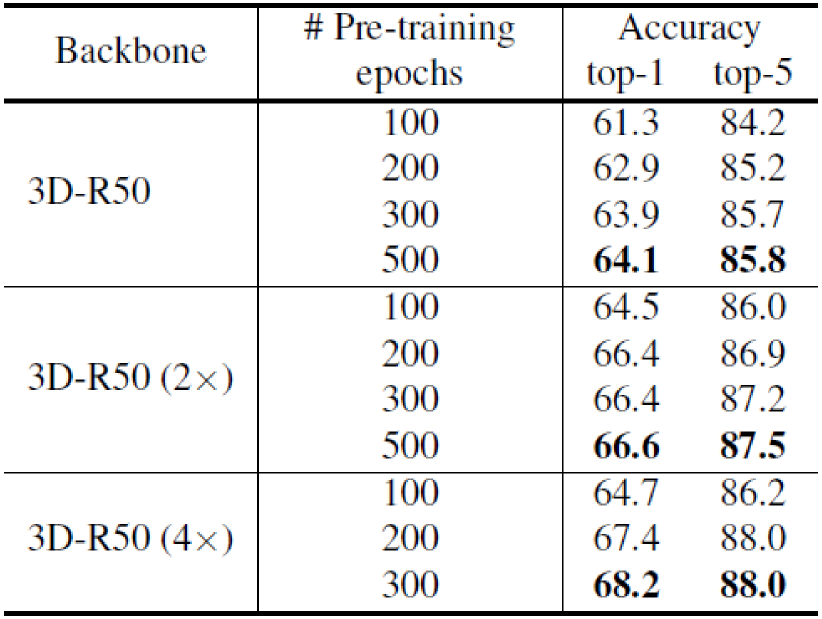
Backbone	Hidden layers	Params	Accuracy	
			top-1	top-5
3D-R50	0	31.9M	52.6	77.5
	1	36.1M	61.3	84.2
	2	40.3M	62.2	84.6
	3	44.5M	**62.3**	**84.7**

Table 5. **Performance of different number of hidden layers in MLP** used in CVRL pre-training (100 epochs). "Params" indicates the total number of parameters in the pre-training network. Results of linear evaluation on top of the same backbone are reported.



Backbone	Batch size	Accuracy	
		top-1	top-5
3D-R50	256	60.4	83.1
	512	61.1	83.7
	1024	**61.3**	**84.2**
	2048	58.4	82.5

Table 7. **Performance of different batch sizes** used in CVRL pre-training (100 epochs). Linear evaluation results are reported.



Backbone	# Pre-training epochs	Accuracy	
		top-1	top-5
3D-R50	100	61.3	84.2
	200	62.9	85.2
	300	63.9	85.7
	500	**64.1**	**85.8**
3D-R50 (2×)	100	64.5	86.0
	200	66.4	86.9
	300	66.4	87.2
	500	**66.6**	**87.5**
3D-R50 (4×)	100	64.7	86.2
	200	67.4	88.0
	300	**68.2**	**88.0**

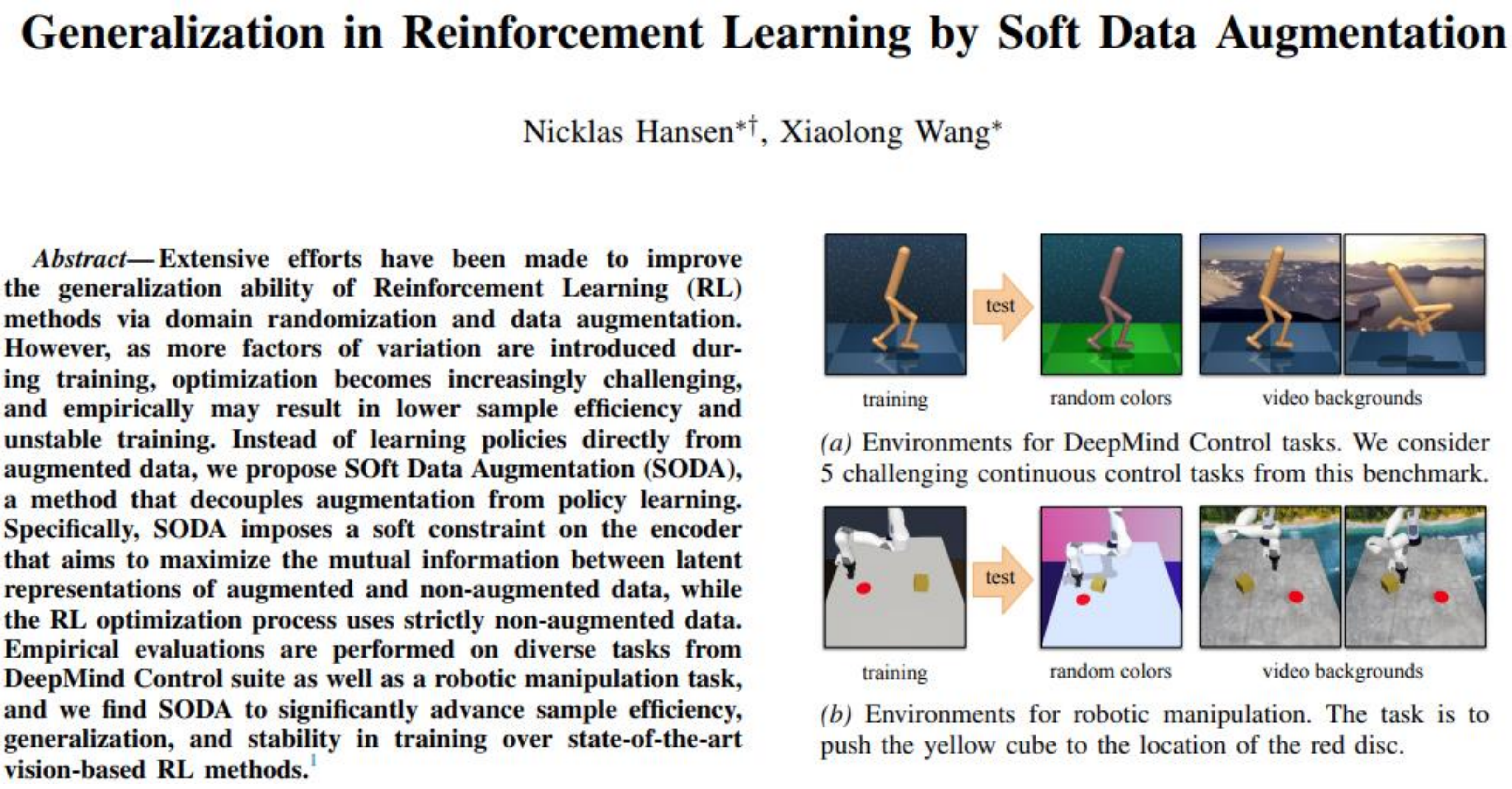
Table 9. **Performance of different pre-training epochs.** The performance starts to saturate after 300 epochs.

Applications of Self-Supervised Learning

Reinforcement Learning

❖ Generalization in Reinforcement Learning by Soft Data Augmentation (IEEE ICRA, 2021)

- California 대학에서 연구하였고 2022년 03월 03일 기준 약 20회 인용



Generalization in Reinforcement Learning by Soft Data Augmentation

Nicklas Hansen*†, Xiaolong Wang*

Abstract—Extensive efforts have been made to improve the generalization ability of Reinforcement Learning (RL) methods via domain randomization and data augmentation. However, as more factors of variation are introduced during training, optimization becomes increasingly challenging, and empirically may result in lower sample efficiency and unstable training. Instead of learning policies directly from augmented data, we propose SOft Data Augmentation (SODA), a method that decouples augmentation from policy learning. Specifically, SODA imposes a soft constraint on the encoder that aims to maximize the mutual information between latent representations of augmented and non-augmented data, while the RL optimization process uses strictly non-augmented data. Empirical evaluations are performed on diverse tasks from DeepMind Control suite as well as a robotic manipulation task, and we find SODA to significantly advance sample efficiency, generalization, and stability in training over state-of-the-art vision-based RL methods.[1]

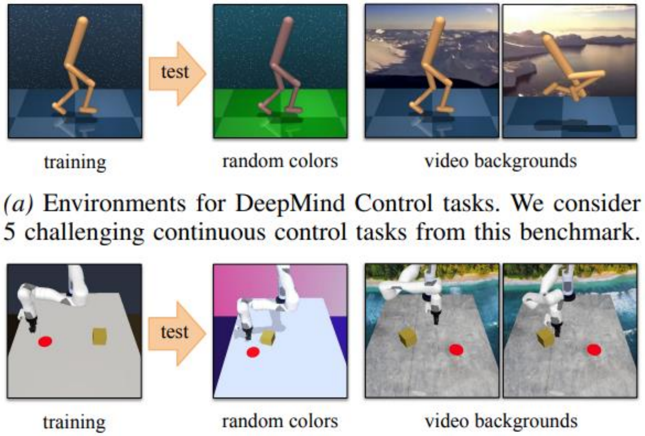
(a) Environments for DeepMind Control tasks. We consider 5 challenging continuous control tasks from this benchmark.

(b) Environments for robotic manipulation. The task is to push the yellow cube to the location of the red disc.

Applications of Self-Supervised Learning

Reinforcement Learning

- ❖ Generalization in Reinforcement Learning by Soft Data Augmentation (IEEE ICRA, 2021)
  - Off-Policy 강화학습은 **샘플 효율성, 학습 안정성, 일반화 능력이 떨어지는** 고질적인 문제 발생이 빈번함
  - 특히, 학습된 에이전트가 한번도 보지 못한 환경에서도 잘 작동하기 위해 데이터 증강 기법을 적용해야 함(**Figure 1, Figure 4**)
  - 본 연구에서는 이미지의 표현을 학습하는 **BYOL 방법**을 적용하여 데이터를 요약하는 **합성곱 인코더를 효과적으로 학습**하는 방법을 제안



(*a*) Environments for DeepMind Control tasks. We consider 5 challenging continuous control tasks from this benchmark.

(*b*) Environments for robotic manipulation. The task is to push the yellow cube to the location of the red disc.

Fig. 1. **Generalization in RL.** Agents are trained in a fixed environment (denoted the *training* environment) and we measure generalization to unseen environments with (i) *random colors* and (ii) *video backgrounds*. To simulate real-world deployment, we additionally randomize camera, lighting, and texture during evaluation in the robotic manipulation task.

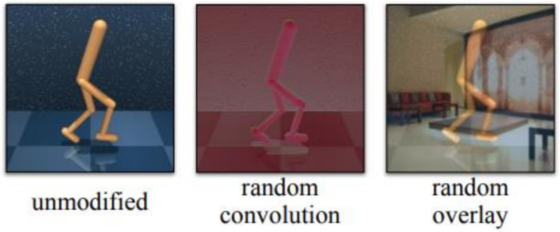


Fig. 4. **Data augmentation.** We consider the following two data augmentations: *random convolution* (as proposed by [9], [16]) and *random overlay* (novel).

Applications of Self-Supervised Learning

Reinforcement Learning

- **SOft Data Augmentation (SODA)**
  - 상태(데이터) 표현 학습 – 공유되는 합성곱 신경망 인코더 $f_\theta(\bullet)$ 학습
    - ✓ 합성곱 신경망 인코더 $f_\theta(\bullet)$를 학습하기 위해 두 단계로 나뉨
    - ✓ Step 1: 증강 기법을 적용하지 않은 상태는 강화학습 적용
    - ✓ Step 2: 증강 기법을 적용한 상태는 BYOL 방법 적용

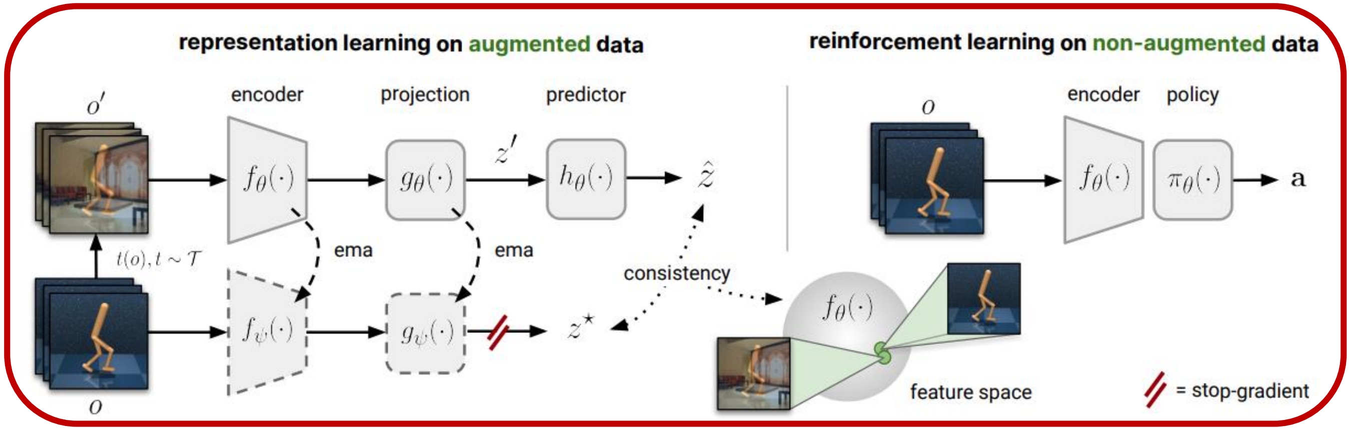


Fig. 2. **SODA architecture.** *Left:* an observation o is augmented to produce a view o', which is then encoded and projected into $z' = g_\theta(f_\theta(o'))$. Likewise, o is encoded by f_ψ and projected by g_ψ to produce features $z^\star$. The SODA objective is then to predict $z^\star$ from z' by h_θ formulated as a consistency loss. *Right:* Reinforcement Learning in SODA. The RL task remains unchanged and is trained directly on the non-augmented observations o. *ema* denotes an exponential moving average.

Applications of Self-Supervised Learning

Reinforcement Learning

❖ SOft Data Augmentation (SODA)

- 상태(데이터) 표현 학습 – 공유되는 합성곱 신경망 인코더 $f_\theta(\bullet)$ 학습
  - ✓ Step 1: 증강 기법을 적용하지 않은 상태는 강화학습 적용
  - ✓ Soft Actor-Critic (SAC) 강화학습 알고리즘을 사용하여 인코더와 강화학습 에이전트 정책(Policy) 학습

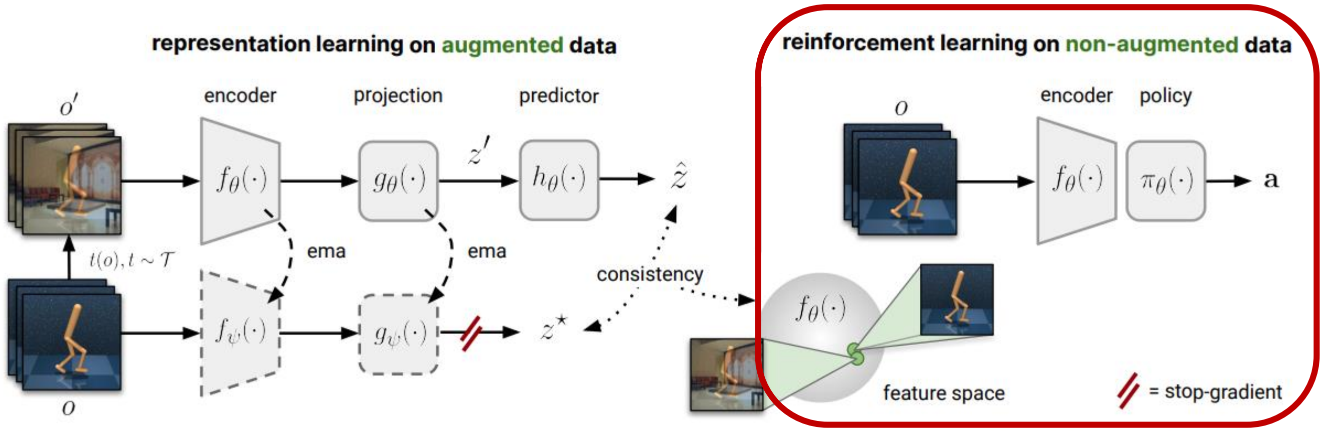


Fig. 2. **SODA architecture.** *Left:* an observation o is augmented to produce a view o', which is then encoded and projected into $z' = g_\theta(f_\theta(o'))$. Likewise, o is encoded by f_ψ and projected by g_ψ to produce features $z^\star$. The SODA objective is then to predict $z^\star$ from z' by h_θ formulated as a consistency loss. *Right:* Reinforcement Learning in SODA. The RL task remains unchanged and is trained directly on the non-augmented observations o. *ema* denotes an exponential moving average.

Applications of Self-Supervised Learning

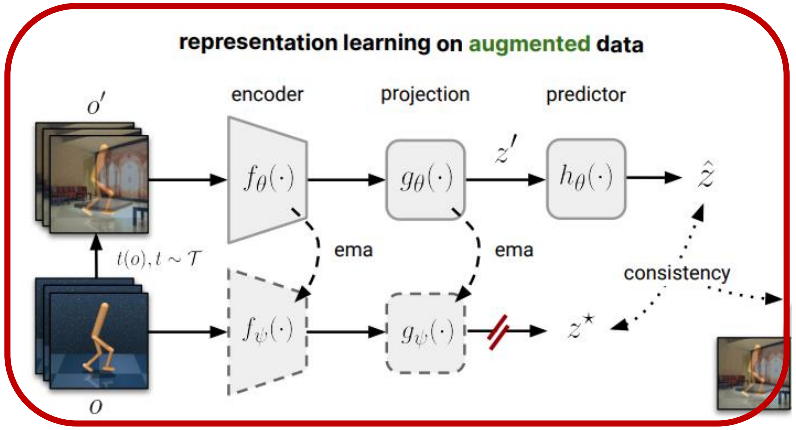
Reinforcement Learning

❖ SOft Data Augmentation (SODA)

- 상태(데이터) 표현 학습 – 공유되는 합성곱 신경망 인코더 $f_\theta(\bullet)$ 학습
  - ✓ Step 2: 증강 기법을 적용한 상태는 BYOL 방법 적용
  - ✓ 원본 상태에 증강 기법을 적용하여 Positive Pair 정의
  - ✓ L2 Loss Function

$$\mathcal{L}_{\text{SODA}}(\hat{z}, z^*; \theta) = \mathrm{E}_{t\sim\mathcal{T}}\left[\left|\left|\hat{z_o} - z_o^*\right|\right|_2^2\right]$$

$$where\ \hat{z} \triangleq h_\theta(z'), \hat{z_o} \triangleq \frac{\hat{z}}{||\hat{z}||_2}, z_o^* \triangleq \frac{z^*}{||z^*||_2}$$



Momentum Update
(Target Network)

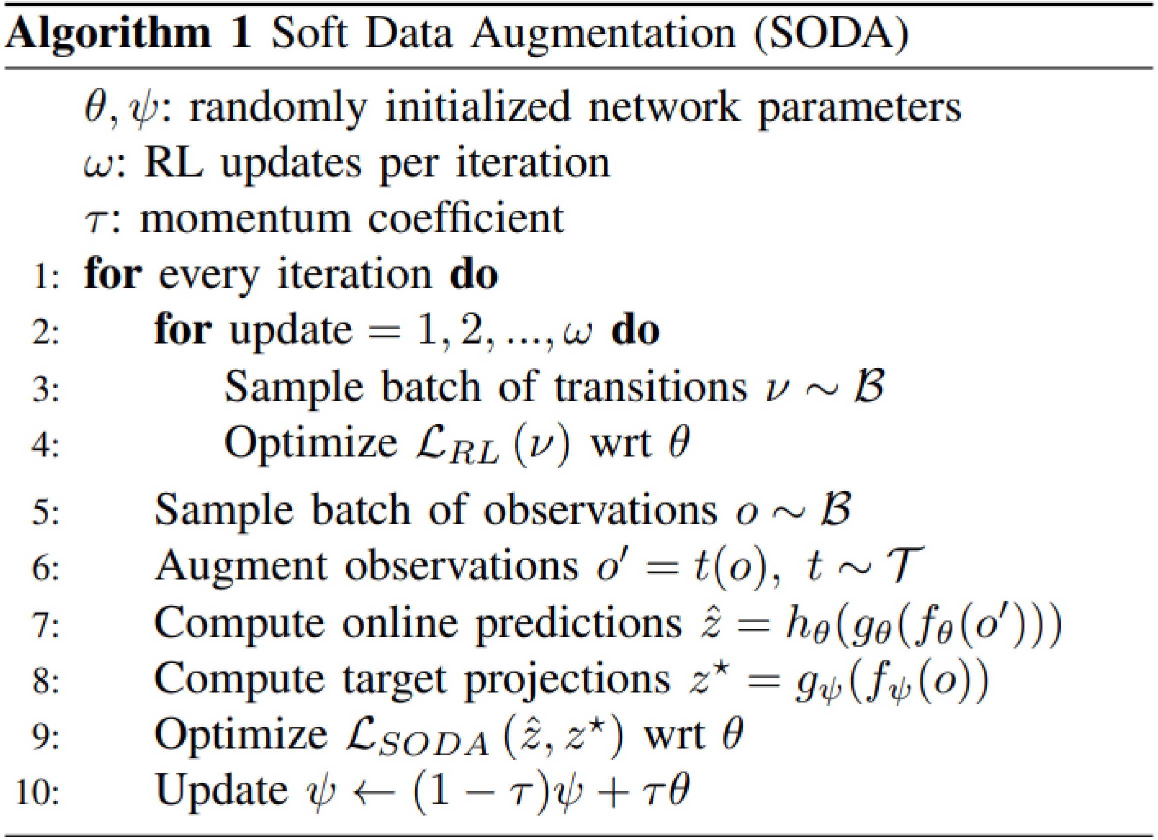
$$\varphi_{target} \leftarrow \tau\varphi_{target} + (1-\tau)\theta_{online}$$

Fig. 2. **SODA architecture.** *Left:* an observation o is augmented to produce a view o', which is then encoded and projected into $z' = g_\theta(f_\theta(o'))$. Likewise, o is encoded by f_ψ and projected by g_ψ to produce features $z^\star$. The SODA objective is then to predict $z^\star$ from z' by h_θ formulated as a consistency loss. *Right:* Reinforcement Learning in SODA. The RL task remains unchanged and is trained directly on the non-augmented observations o. *ema* denotes an exponential moving average.

Applications of Self-Supervised Learning

Reinforcement Learning

- ❖ **SO**ft **D**ata **A**ugmentation (SODA)
  - 상태(데이터) 표현 학습 – 공유되는 합성곱 신경망 인코더 $f_\theta(\bullet)$ 학습
    - ✓ Pseudo Code



Algorithm 1 Soft Data Augmentation (SODA)

```
$\theta, \psi$: randomly initialized network parameters
$\omega$: RL updates per iteration
$\tau$: momentum coefficient
for every iteration do
    for update $= 1, 2, ..., \omega$ do
        Sample batch of transitions $\nu \sim \mathcal{B}$
        Optimize $\mathcal{L}_{RL}(\nu)$ wrt $\theta$
    Sample batch of observations $o \sim \mathcal{B}$
    Augment observations $o' = t(o),\ t \sim \mathcal{T}$
    Compute online predictions $\hat{z} = h_\theta(g_\theta(f_\theta(o')))$
    Compute target projections $z^\star = g_\psi(f_\psi(o))$
    Optimize $\mathcal{L}_{SODA}(\hat{z}, z^\star)$ wrt $\theta$
    Update $\psi \leftarrow (1 - \tau)\psi + \tau\theta$
```

Applications of Self-Supervised Learning

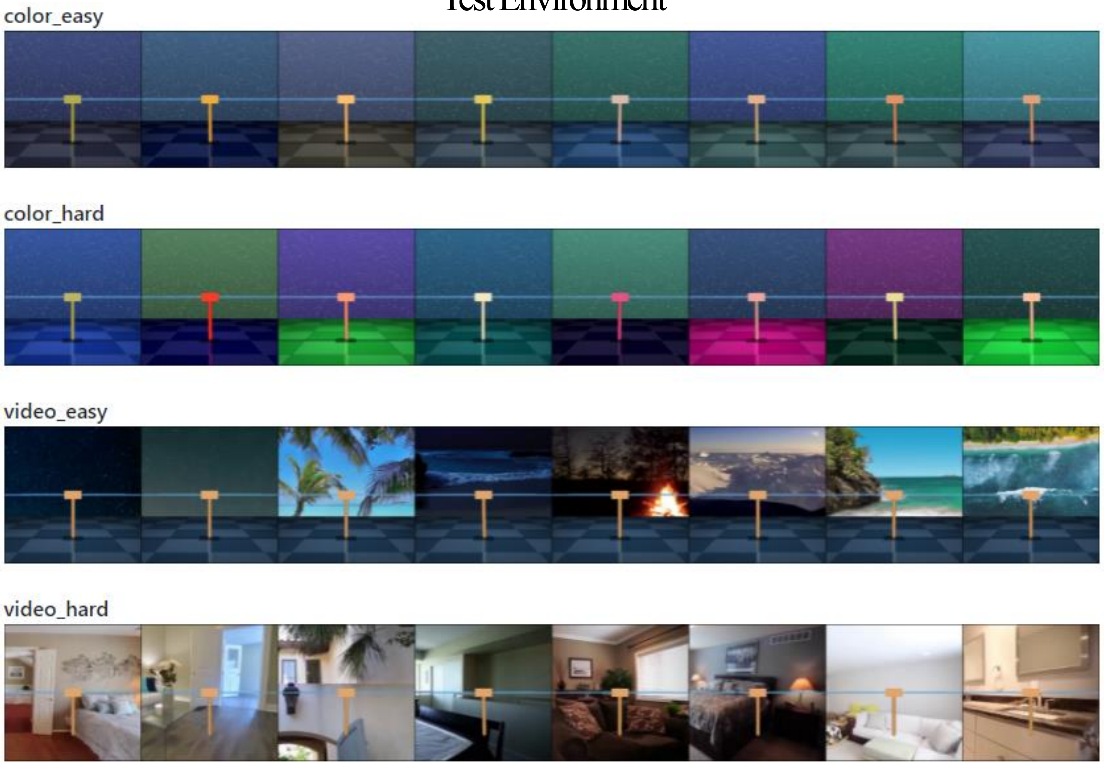
Reinforcement Learning

❖ Experiments – 실험 환경

- DeepMind Control Suite Environment (로봇 제어 작업)
- 학습에 사용되는 환경과 테스트에 사용되는 환경이 다름
- 점수로 샘플 효율성, 학습 안정성, 일반화 성능 향상 판단

Training Environment

Test Environment

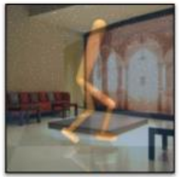


Applications of Self-Supervised Learning

Reinforcement Learning

unmodified

random convolution



random overlay

- ❖ Experiments – 학습 안정성, 샘플 효율성 판단
  - 5개의 환경에 적용
  - 고정된 학습 시간 동안 점수가 상승하는지 그림으로 확인
  - 제안 방법론에 Random Convolution 증강 기법을 적용했을 때 결과

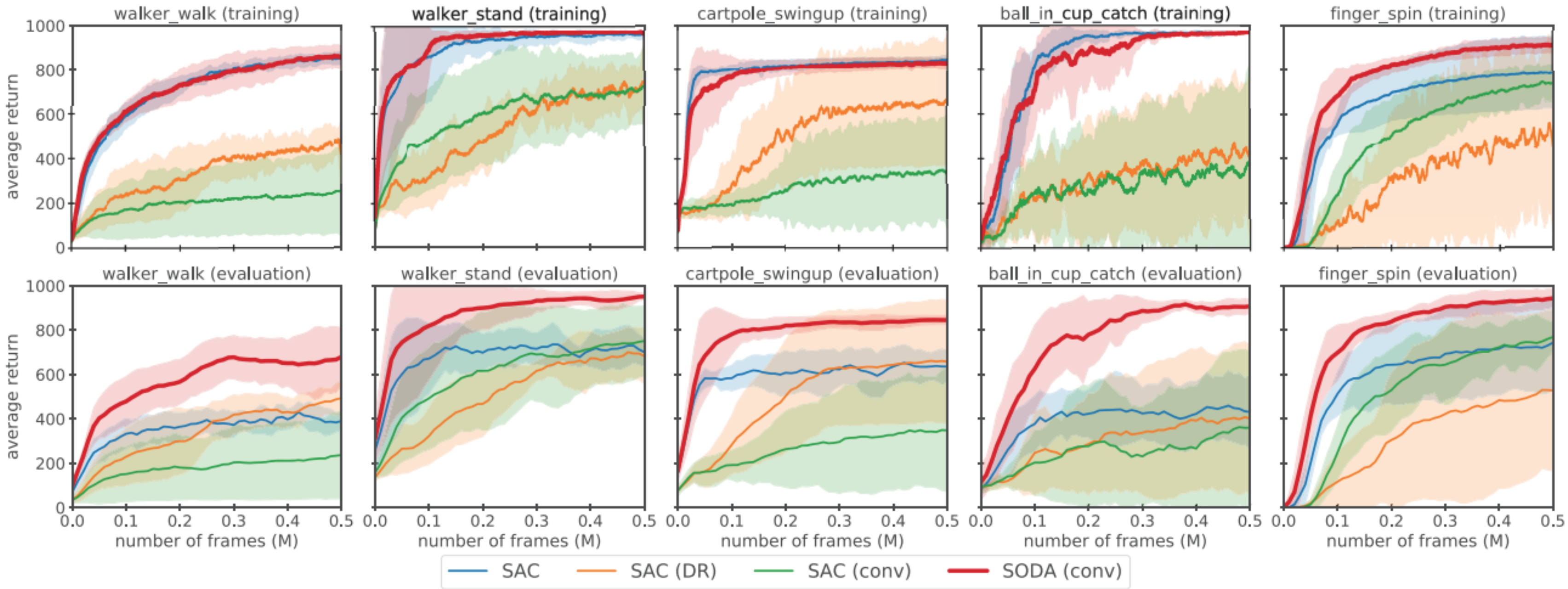
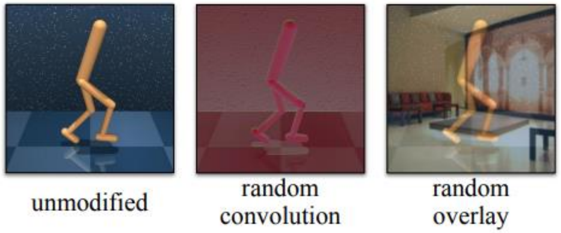


Fig. 3. **Random convolution.** *Top:* average return on the training environment during training. *Bottom:* periodic evaluation of generalization ability measured by average return on the *random color* environment. SODA exhibits sample efficiency and convergence similar to SAC but improves generalization significantly. Average of 5 runs, shaded area is std. deviation.

Applications of Self-Supervised Learning

Reinforcement Learning


unmodified

random convolution

random overlay

- ❖ Experiments – 학습 안정성, 샘플 효율성 판단
  - 5개의 환경에 적용
  - 고정된 학습 시간 동안 점수가 상승하는지 그림으로 확인
  - 제안 방법론에 Random Overlay 증강 기법을 적용했을 때 결과

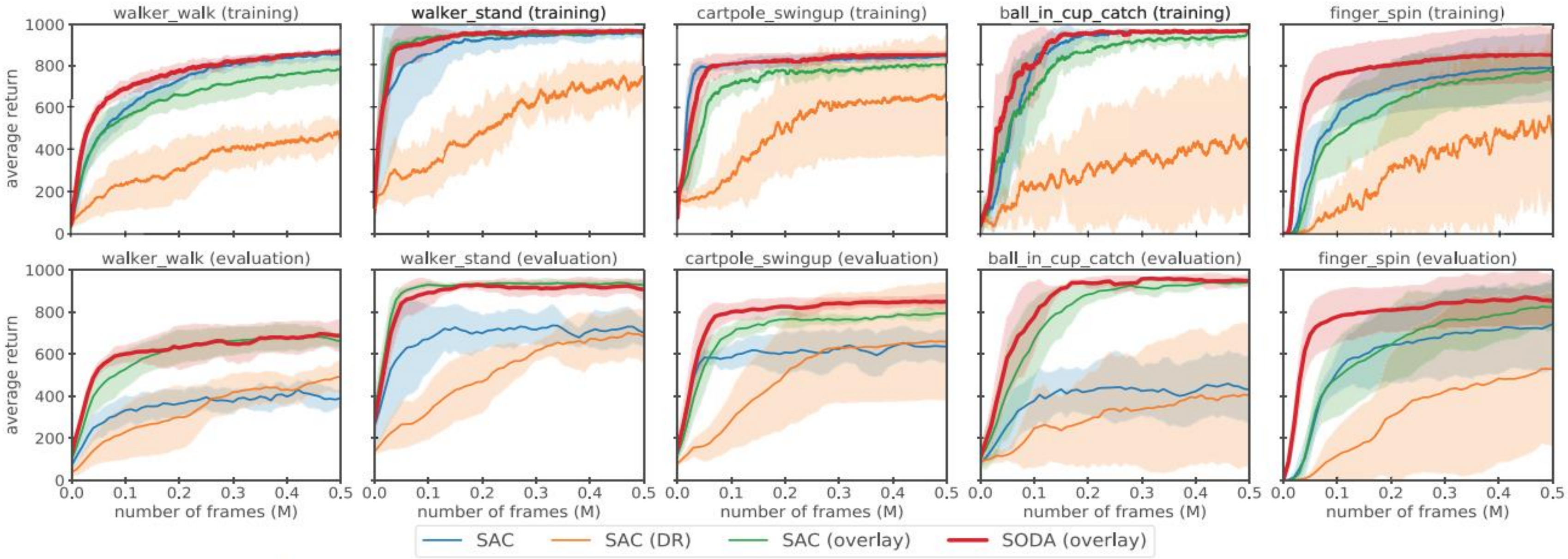


Fig. 6. **Random overlay.** *Top:* average return on the training environment during training. *Bottom:* periodic evaluation of generalization ability measured by average return on the *random color* environment. SODA offers better sample-efficiency than the novel *SAC (overlay)* baseline and similar generalization to *SODA (conv)* even though there is minimal visual similarity between random overlays and the random color environment. Average of 5 runs, shaded area is std. deviation.

Applications of Self-Supervised Learning

Reinforcement Learning

- ❖ Experiments – 일반화 성능 판단
  - 학습된 에이전트가 한번도 보지 못한 환경 구성(Video Backgrounds, Random Colors)
  - 테스트 환경에서 5번 반복 실험하여 얻은 점수의 평균으로 평가

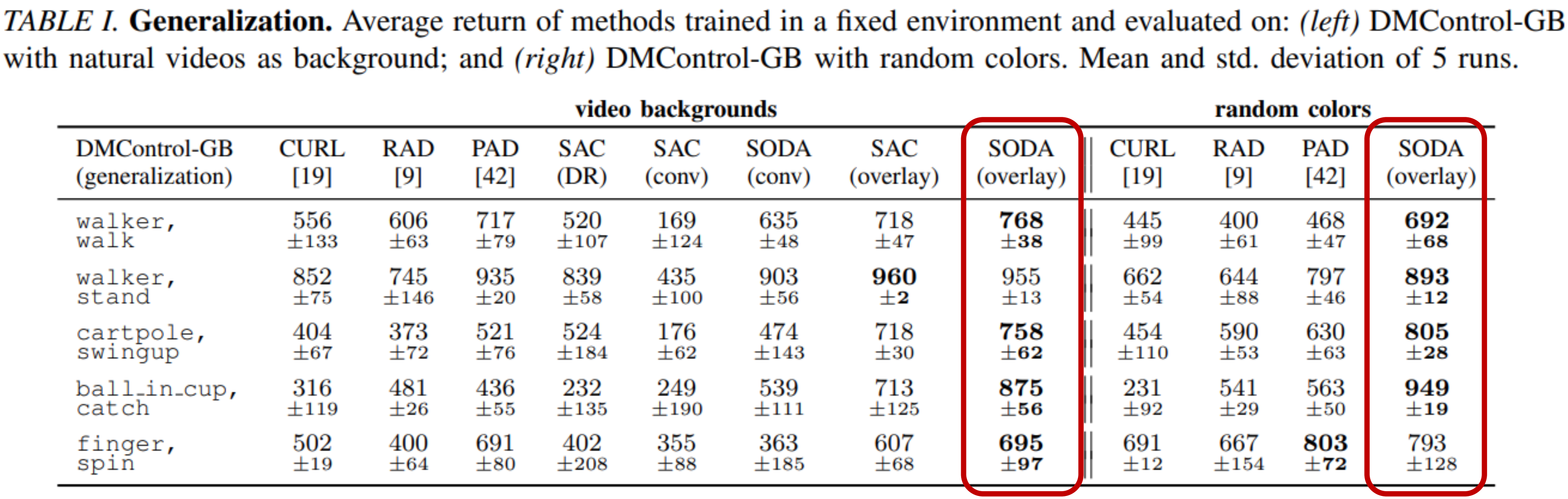


TABLE I. **Generalization.** Average return of methods trained in a fixed environment and evaluated on: *(left)* DMControl-GB with natural videos as background; and *(right)* DMControl-GB with random colors. Mean and std. deviation of 5 runs.

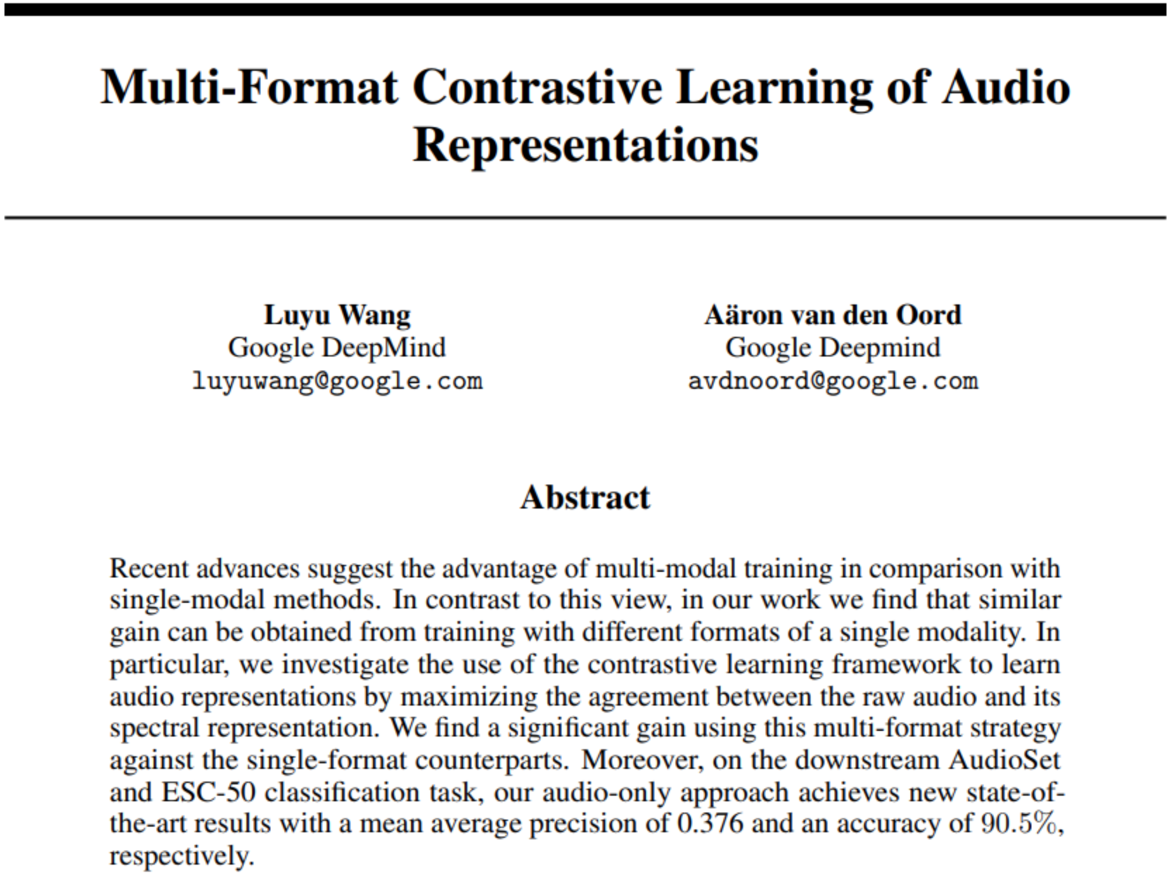
				video backgrounds						**random colors**		
DMControl-GB (generalization)	CURL [19]	RAD [9]	PAD [42]	SAC (DR)	SAC (conv)	SODA (conv)	SAC (overlay)	SODA (overlay)	CURL [19]	RAD [9]	PAD [42]	SODA (overlay)
walker, walk	556 ±133	606 ±63	717 ±79	520 ±107	169 ±124	635 ±48	718 ±47	**768 ±38**	445 ±99	400 ±61	468 ±47	**692 ±68**
walker, stand	852 ±75	745 ±146	935 ±20	839 ±58	435 ±100	903 ±56	**960 ±2**	955 ±13	662 ±54	644 ±88	797 ±46	**893 ±12**
cartpole, swingup	404 ±67	373 ±72	521 ±76	524 ±184	176 ±62	474 ±143	718 ±30	**758 ±62**	454 ±110	590 ±53	630 ±63	**805 ±28**
ball_in_cup, catch	316 ±119	481 ±26	436 ±55	232 ±135	249 ±190	539 ±111	713 ±125	**875 ±56**	231 ±92	541 ±29	563 ±50	**949 ±19**
finger, spin	502 ±19	400 ±64	691 ±80	402 ±208	355 ±88	363 ±185	607 ±68	**695 ±97**	691 ±12	667 ±154	**803 ±72**	793 ±128

Applications of Self-Supervised Learning

Audio

❖ Multi-Format Contrastive Learning of Audio Representations (arXiv, 2021)

- Google DeepMind에서 연구하였고 2022년 03월 03일 기준 약 14회 인용



Multi-Format Contrastive Learning of Audio Representations

Luyu Wang
Google DeepMind
luyuwang@google.com

Äàron van den Oord
Google Deepmind
avdnoord@google.com

Abstract

Recent advances suggest the advantage of multi-modal training in comparison with single-modal methods. In contrast to this view, in our work we find that similar gain can be obtained from training with different formats of a single modality. In particular, we investigate the use of the contrastive learning framework to learn audio representations by maximizing the agreement between the raw audio and its spectral representation. We find a significant gain using this multi-format strategy against the single-format counterparts. Moreover, on the downstream AudioSet and ESC-50 classification task, our audio-only approach achieves new state-of-the-art results with a mean average precision of 0.376 and an accuracy of 90.5%, respectively.

Applications of Self-Supervised Learning

Audio

- Multi-Format Contrastive Learning of Audio Representations (arXiv, 2021)
  - 최근 연구에서는 **Multi-Modal** 학습이 Single-Modal 방법들보다 효과적임을 증명
  - 오디오 데이터는 Waveform, Spectrogram으로 표현이 가능
  - **Single Modality 데이터**(Waveform)에 **서로 다른 포맷의 데이터**(Waveform, Spectrogram)를 추출하여 오디오의 표현을 학습하는 **Multi-Format 대조 학습** 제안

Applications of Self-Supervised Learning

Audio

❖ Multi-Format Contrastive Audio Learning

- 오디오 표현 학습을 위해 SimCLR 적용 – Positive와 Negative Examples 정의
  - ✓ 한 오디오 데이터 내 서로 다른 시점의 두 Waveform 선택
  - ✓ 하나의 Waveform은 그대로 사용하고 다른 하나는 Spectrogram으로 변형하여 사용(Multi-Format)
  - ✓ 서로 다른 포맷의 데이터는 Positive Pair, 다른 오디오 데이터에서 추출한 데이터는 Negative Examples

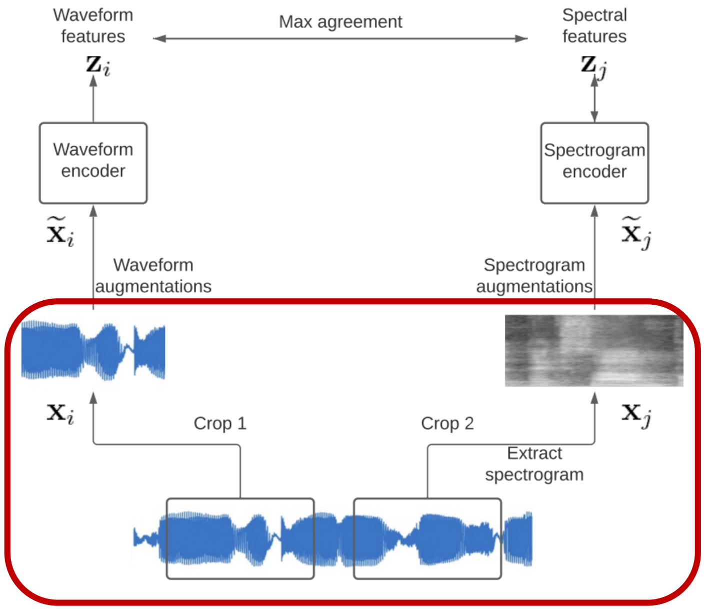


Figure 1: Illustration of the multi-format contrastive audio learning framework.

Applications of Self-Supervised Learning

Audio

❖ Multi-Format Contrastive Audio Learning

- 오디오 표현 학습을 위해 SimCLR 적용 – 데이터 증강 기법
  - ✓ Waveform, Spectrogram 데이터에 알맞은 데이터 증강 기법 적용
  - ✓ Audio Mixing, Time Masking, Frequency Masking, Frequency Shift, . . .

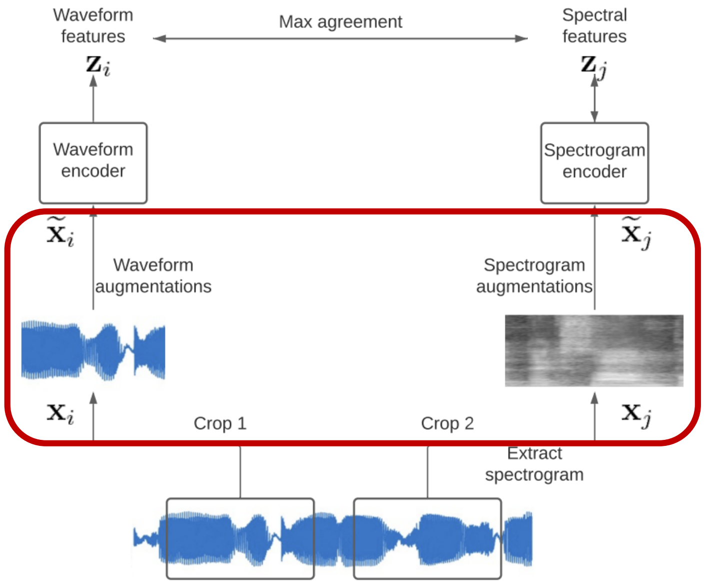


Figure 1: Illustration of the multi-format contrastive audio learning framework.

Applications of Self-Supervised Learning

Audio

- ❖ Multi-Format Contrastive Audio Learning
  - 오디오 표현 학습을 위해 SimCLR 적용 – 데이터 요약
    - ✓ 1차원 데이터인 Waveform은 Res1dNet-31 적용하여 요약
    - ✓ 2차원 데이터인 Spectrogram은 CNN14 적용하여 요약

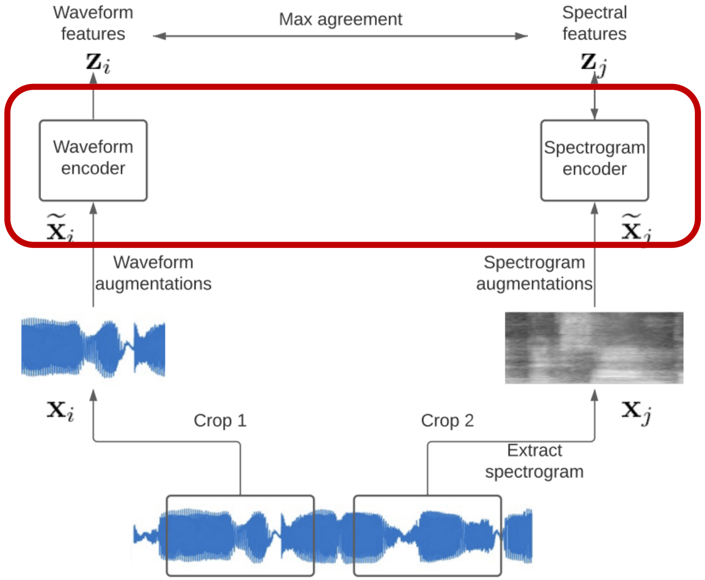


Figure 1: Illustration of the multi-format contrastive audio learning framework.

Applications of Self-Supervised Learning

Audio

- ❖ Multi-Format Contrastive Audio Learning
  - 오디오 표현 학습을 위해 SimCLR 적용 – InfoNCE Loss Function
    - ✓ 각 인코더로부터 요약된 벡터들을 사용하여 대조 학습 수행

$$\mathcal{L}_i = -log \frac{exp\left(\frac{i \cdot i^+}{\tau}\right)}{exp\left(\frac{i \cdot i^+}{\tau}\right) + \sum_{k \notin \{i, i^+\}} exp\left(\frac{i \cdot k}{\tau}\right)}$$

i: Anchor

i^+: Positive

k: Negative Examples

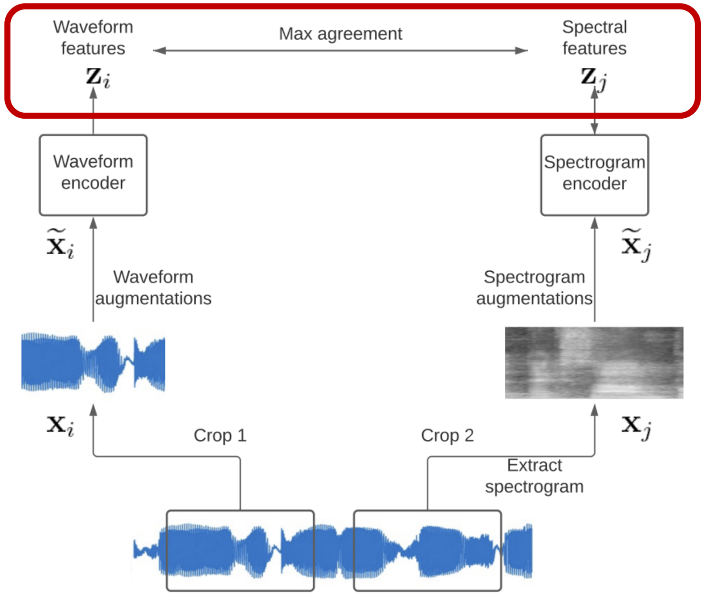


Figure 1: Illustration of the multi-format contrastive audio learning framework.

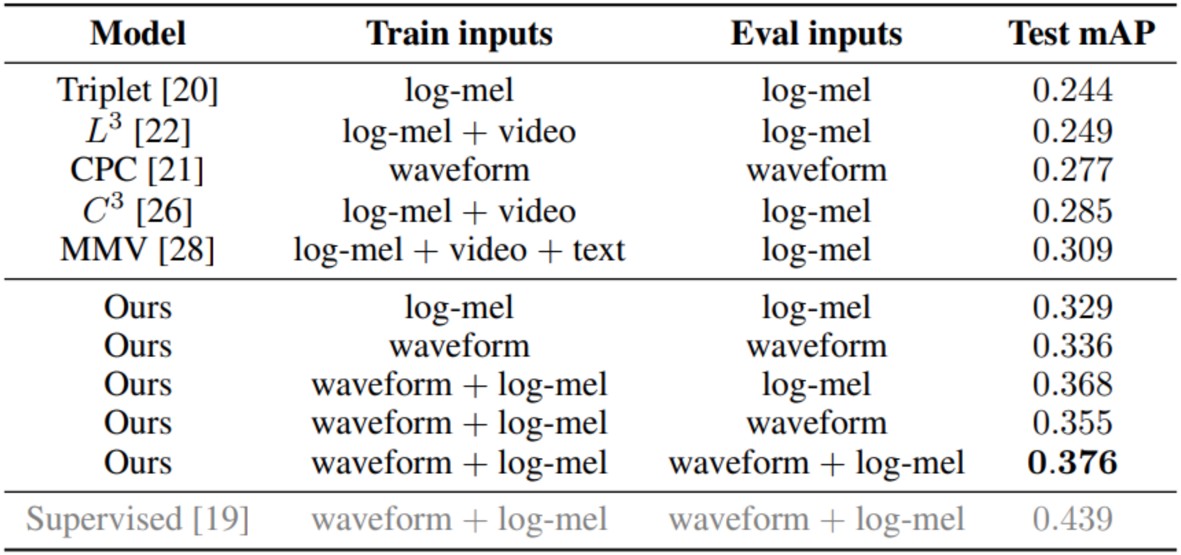
Applications of Self-Supervised Learning

Audio

❖ Experiments – Audio Classification

- AudioSet 사용
- Mean Average Precision로 평가(0.376)

Table 3: Test performance of shallow model classification on AudioSet with fixed representations.



Model	Train inputs	Eval inputs	Test mAP
Triplet [20]	log-mel	log-mel	0.244
L^3 [22]	log-mel + video	log-mel	0.249
CPC [21]	waveform	waveform	0.277
C^3 [26]	log-mel + video	log-mel	0.285
MMV [28]	log-mel + video + text	log-mel	0.309
Ours	log-mel	log-mel	0.329
Ours	waveform	waveform	0.336
Ours	waveform + log-mel	log-mel	0.368
Ours	waveform + log-mel	waveform	0.355
Ours	waveform + log-mel	waveform + log-mel	**0.376**
Supervised [19]	waveform + log-mel	waveform + log-mel	0.439

Applications of Self-Supervised Learning

Audio

❖ Experiments – Audio Classification

- ESC-50 데이터셋 사용
- Accuracy로 평가(90.5%)

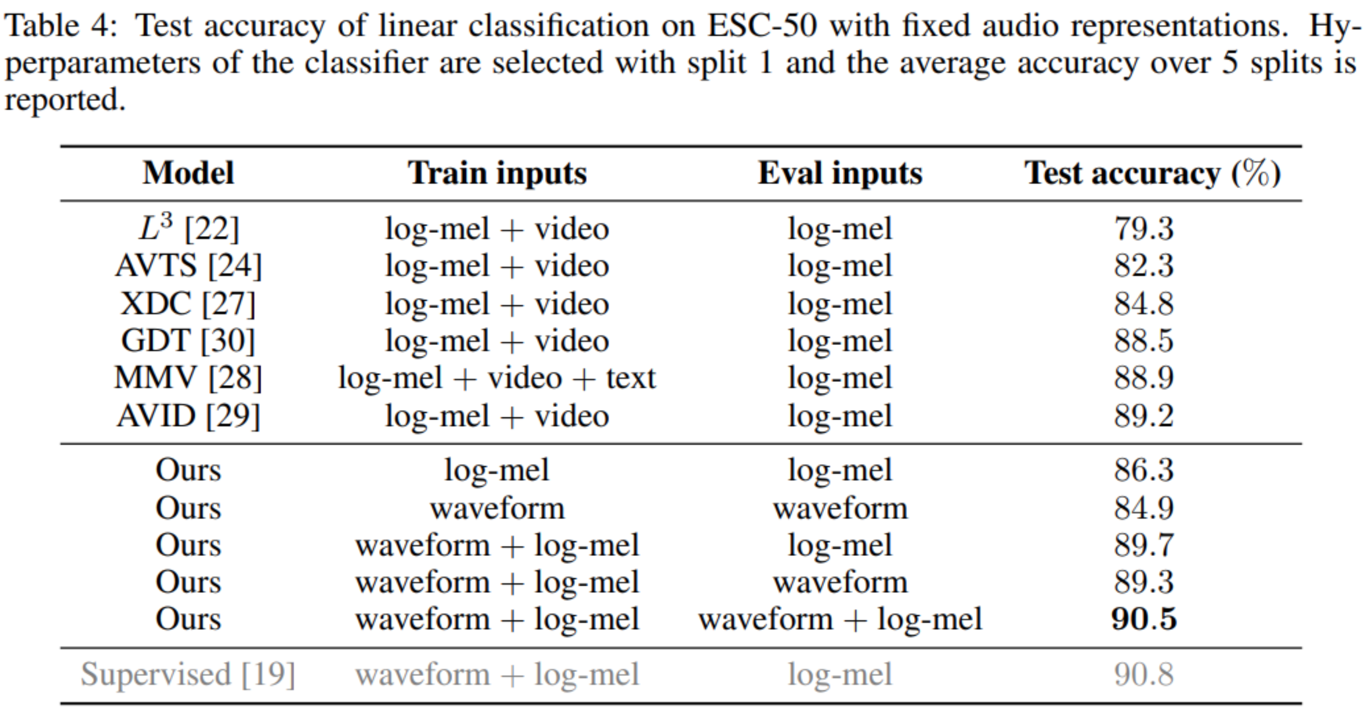


Table 4: Test accuracy of linear classification on ESC-50 with fixed audio representations. Hyperparameters of the classifier are selected with split 1 and the average accuracy over 5 splits is reported.

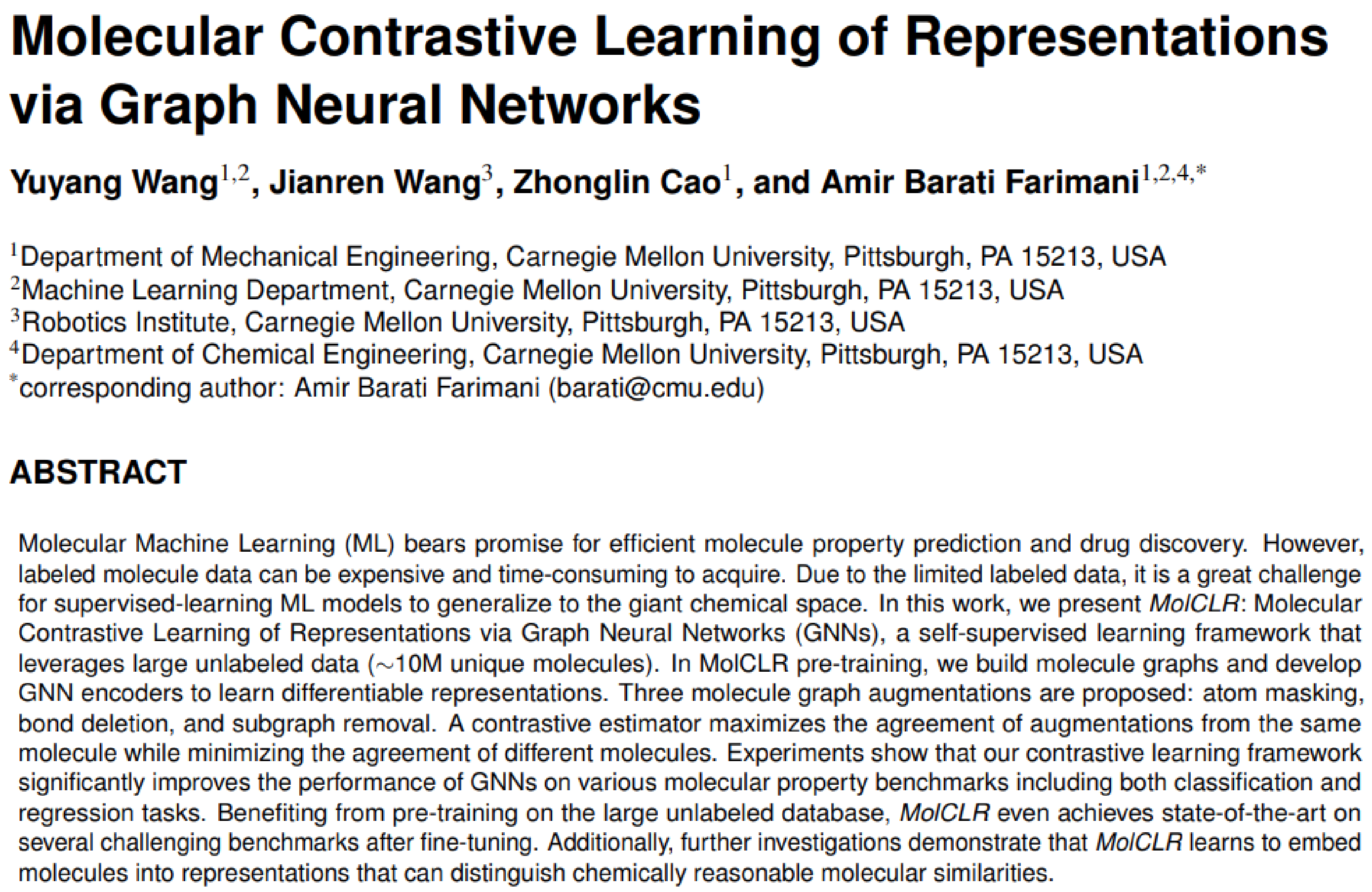
Model	**Train inputs**	**Eval inputs**	**Test accuracy** (%)
L^3 [22]	log-mel + video	log-mel	79.3
AVTS [24]	log-mel + video	log-mel	82.3
XDC [27]	log-mel + video	log-mel	84.8
GDT [30]	log-mel + video	log-mel	88.5
MMV [28]	log-mel + video + text	log-mel	88.9
AVID [29]	log-mel + video	log-mel	89.2
Ours	log-mel	log-mel	86.3
Ours	waveform	waveform	84.9
Ours	waveform + log-mel	log-mel	89.7
Ours	waveform + log-mel	waveform	89.3
Ours	waveform + log-mel	waveform + log-mel	**90.5**
Supervised [19]	waveform + log-mel	log-mel	90.8

Applications of Self-Supervised Learning

Graph

❖ Molecular Contrastive Learning of Representations via Graph Neural Networks (arXiv, 2021)

- Carnegie Mellon 대학에서 연구하였고 2022년 03월 03일 기준 약 17회 인용



Molecular Contrastive Learning of Representations via Graph Neural Networks

Yuyang Wang[1,2], **Jianren Wang**[3], **Zhonglin Cao**[1], **and Amir Barati Farimani**[1,2,4,*]

[1]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[*]corresponding author: Amir Barati Farimani (barati@cmu.edu)

ABSTRACT

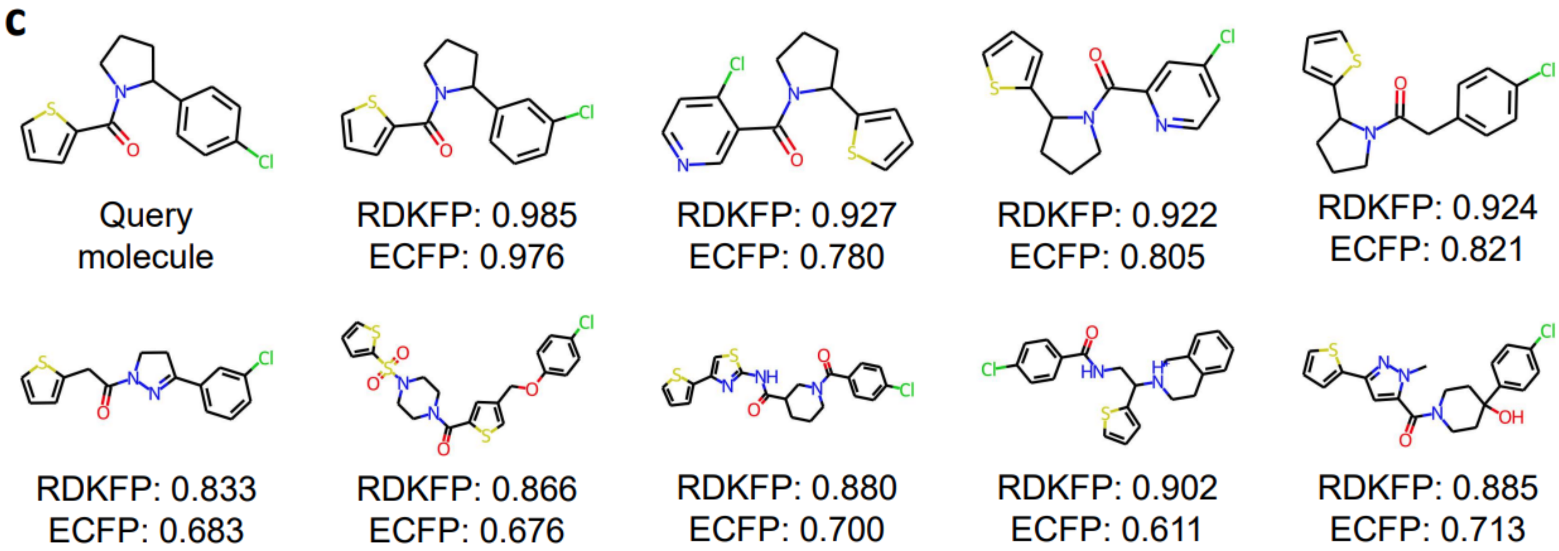
Molecular Machine Learning (ML) bears promise for efficient molecule property prediction and drug discovery. However, labeled molecule data can be expensive and time-consuming to acquire. Due to the limited labeled data, it is a great challenge for supervised-learning ML models to generalize to the giant chemical space. In this work, we present *MolCLR*: Molecular Contrastive Learning of Representations via Graph Neural Networks (GNNs), a self-supervised learning framework that leverages large unlabeled data (~10M unique molecules). In MolCLR pre-training, we build molecule graphs and develop GNN encoders to learn differentiable representations. Three molecule graph augmentations are proposed: atom masking, bond deletion, and subgraph removal. A contrastive estimator maximizes the agreement of augmentations from the same molecule while minimizing the agreement of different molecules. Experiments show that our contrastive learning framework significantly improves the performance of GNNs on various molecular property benchmarks including both classification and regression tasks. Benefiting from pre-training on the large unlabeled database, *MolCLR* even achieves state-of-the-art on several challenging benchmarks after fine-tuning. Additionally, further investigations demonstrate that *MolCLR* learns to embed molecules into representations that can distinguish chemically reasonable molecular similarities.

Applications of Self-Supervised Learning

Graph

❖ Molecular Contrastive Learning of Representations via Graph Neural Networks (arXiv, 2021)

- 화학 분야 중 지도 학습 기반 분자 성질 예측, 신약 발견 분야에서 큰 성공을 거두었음
- 하지만, 레이블이 있는 데이터를 수집하는데 **비용, 시간 소비가 매우 큼**
- 레이블이 없는 데이터를 사용하여 **분자 정보를 세밀하게 표현하는 것**이 핵심
- 본 연구에서는 천만 개의 레이블이 없는 데이터(PubChem)를 사용한 대조 학습 방법을 제안
  - ✓ 특히, 분자 구조를 그래프로 표현하여 **그래프 기반 대조 학습** 제안

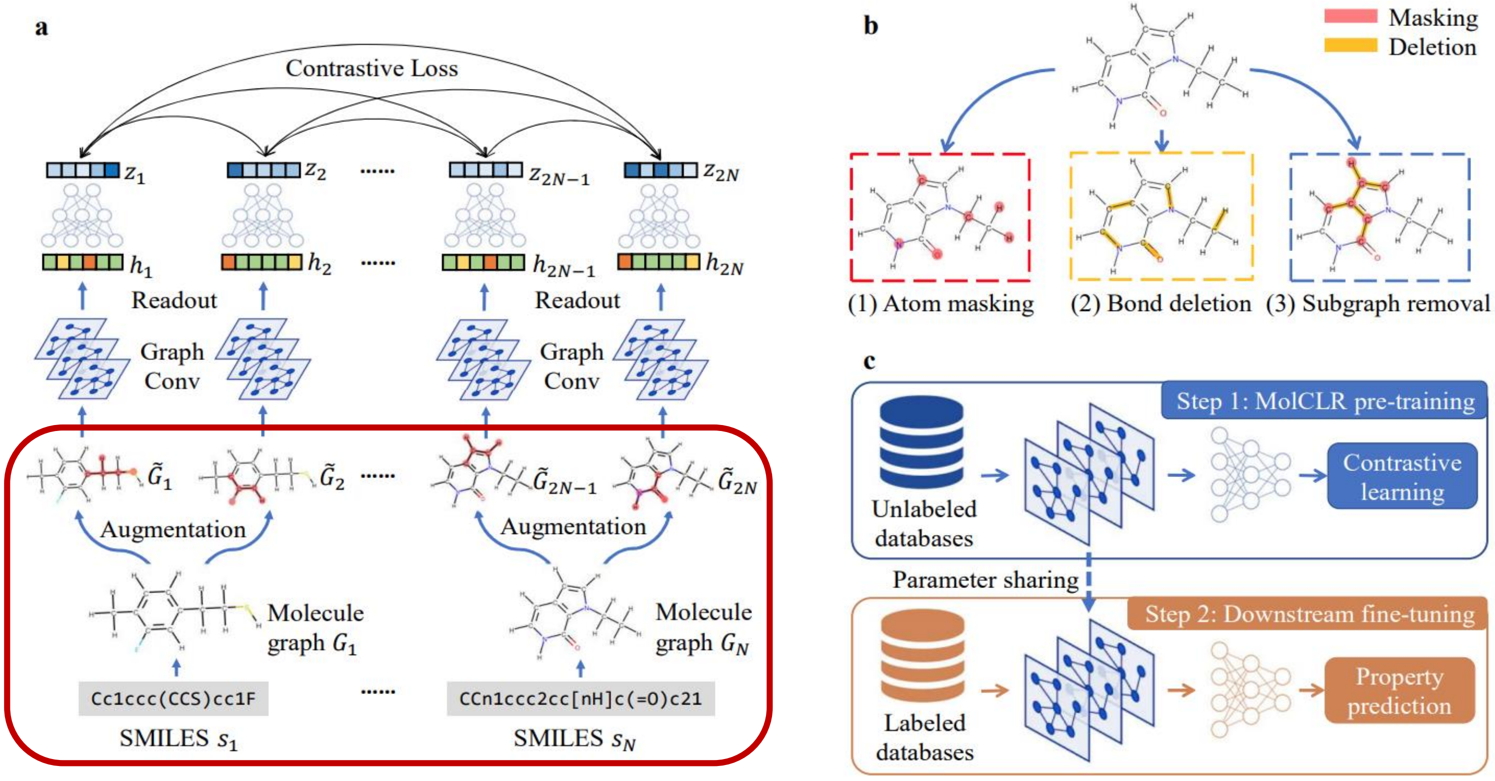


Applications of Self-Supervised Learning

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

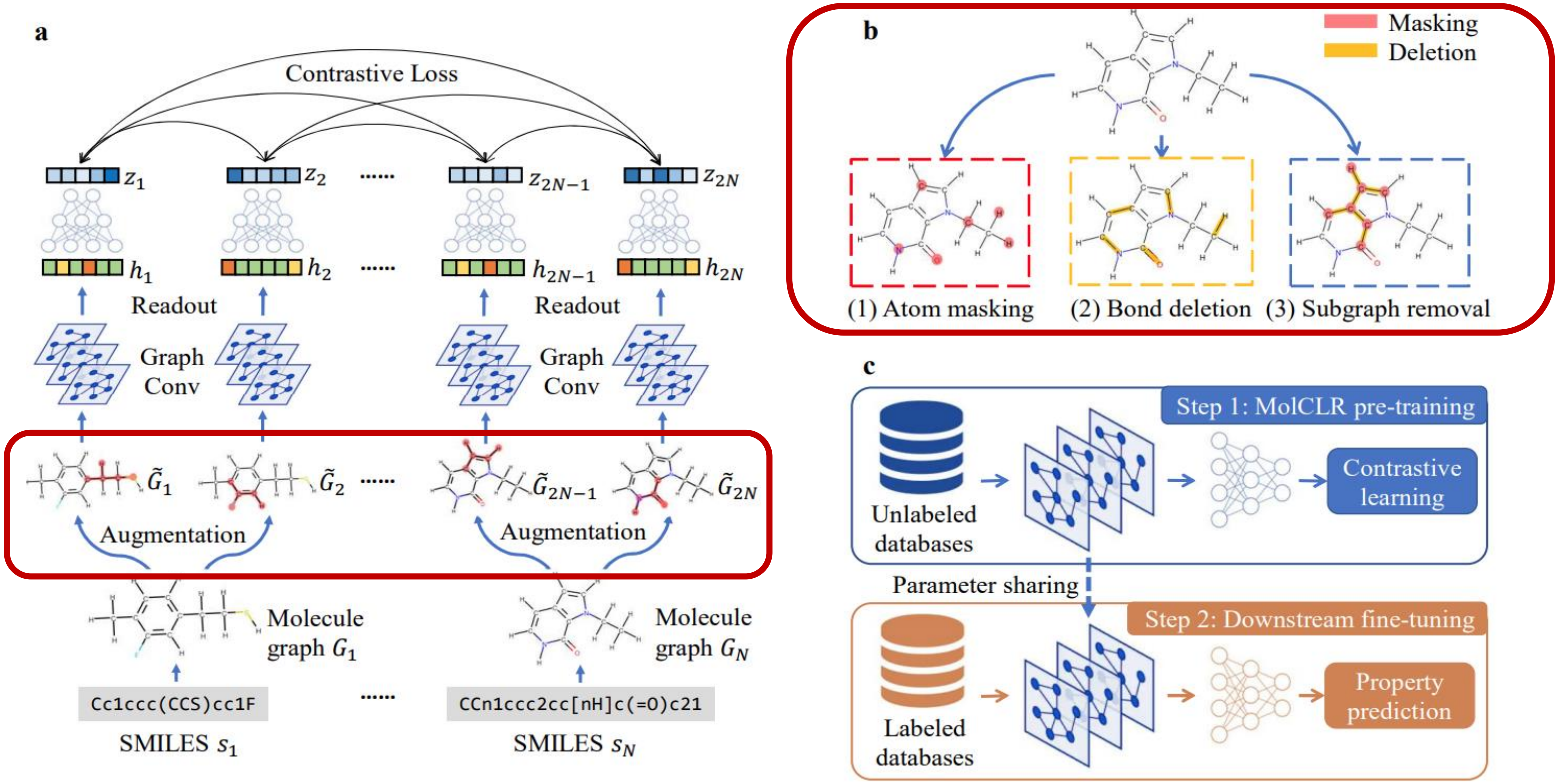
- 분자 그래프 표현 학습을 위해 SimCLR 적용 – Positive Pair와 Negative Examples 정의
  - ✓ 시퀀스 형태로 표현되는 분자(SMILES String)를 그래프 형태로 변환
  - ✓ 한 분자 그래프에서 서로 다른 증강 기법을 적용한 쌍 Positive Pair $(\tilde{G}_1, \tilde{G}_2)$, 이 외에는 Negative Examples



Applications of Self-Supervised Learning

Graph

- ❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)
  - 분자 그래프 표현 학습을 위해 SimCLR 적용 – 데이터 증강 기법
    - ✓ Atom Masking, Bond Deletion, Subgraph Removal 적용
    - ✓ 일정 비율로 원자 정보를 삭제, 원자 간 결합 정보를 삭제, 원자와 결합 정보 일부를 통으로 삭제

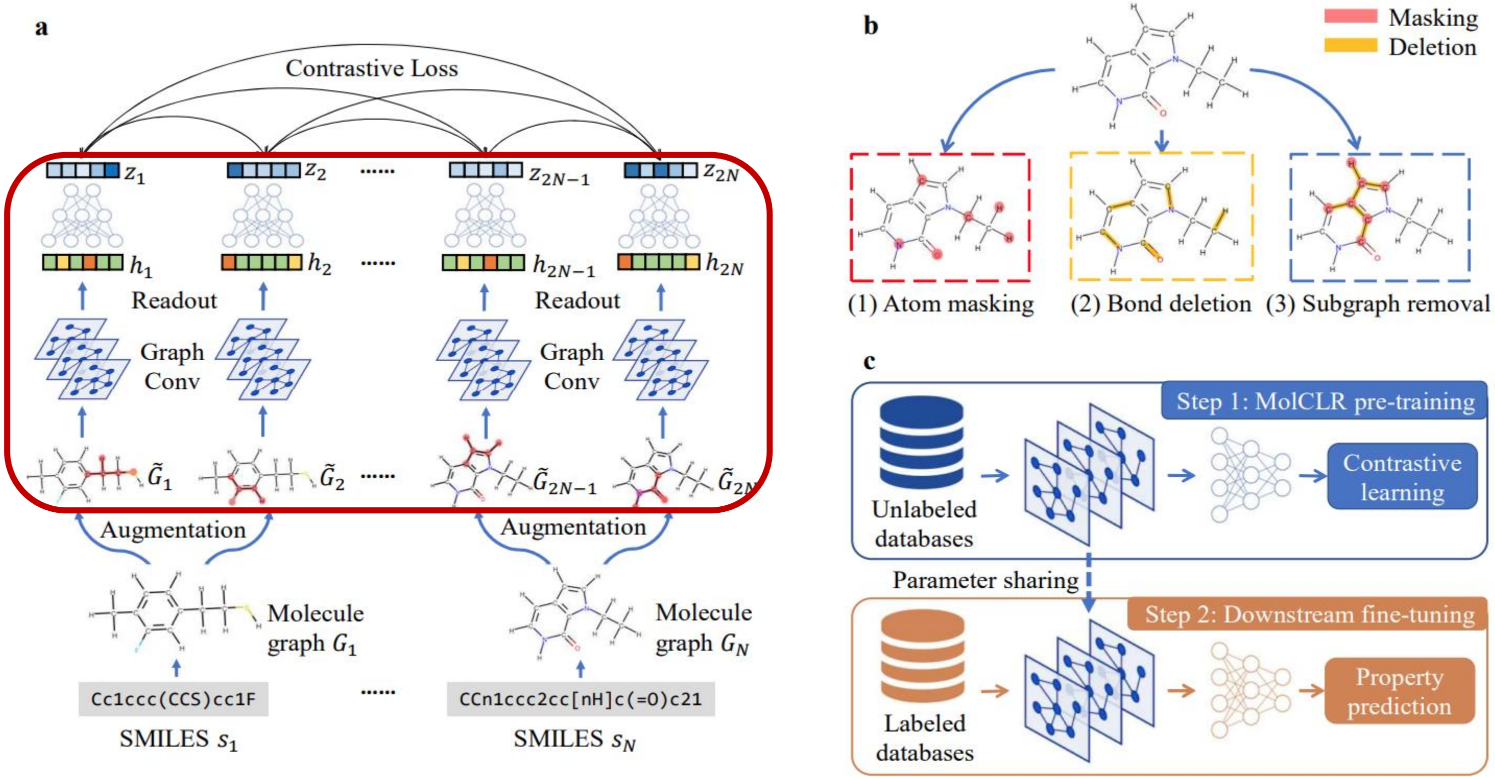


Applications of Self-Supervised Learning

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – 분자 그래프 요약
    - ✓ Graph Convolutional Network (GCN), Graph Isomorphism Network (GIN) 인코더 적용
    - ✓ 그래프 인코더로 요약된 특징을 MLP로 한번 더 요약



Applications of Self-Supervised Learning

$$\mathcal{L}_i = -log\frac{exp\left(\frac{i \cdot i^{+}}{\tau}\right)}{exp\left(\frac{i \cdot i^{+}}{\tau}\right) + \sum_{k \notin \{i, i^{+}\}} exp\left(\frac{i \cdot k}{\tau}\right)}$$

i: Anchor

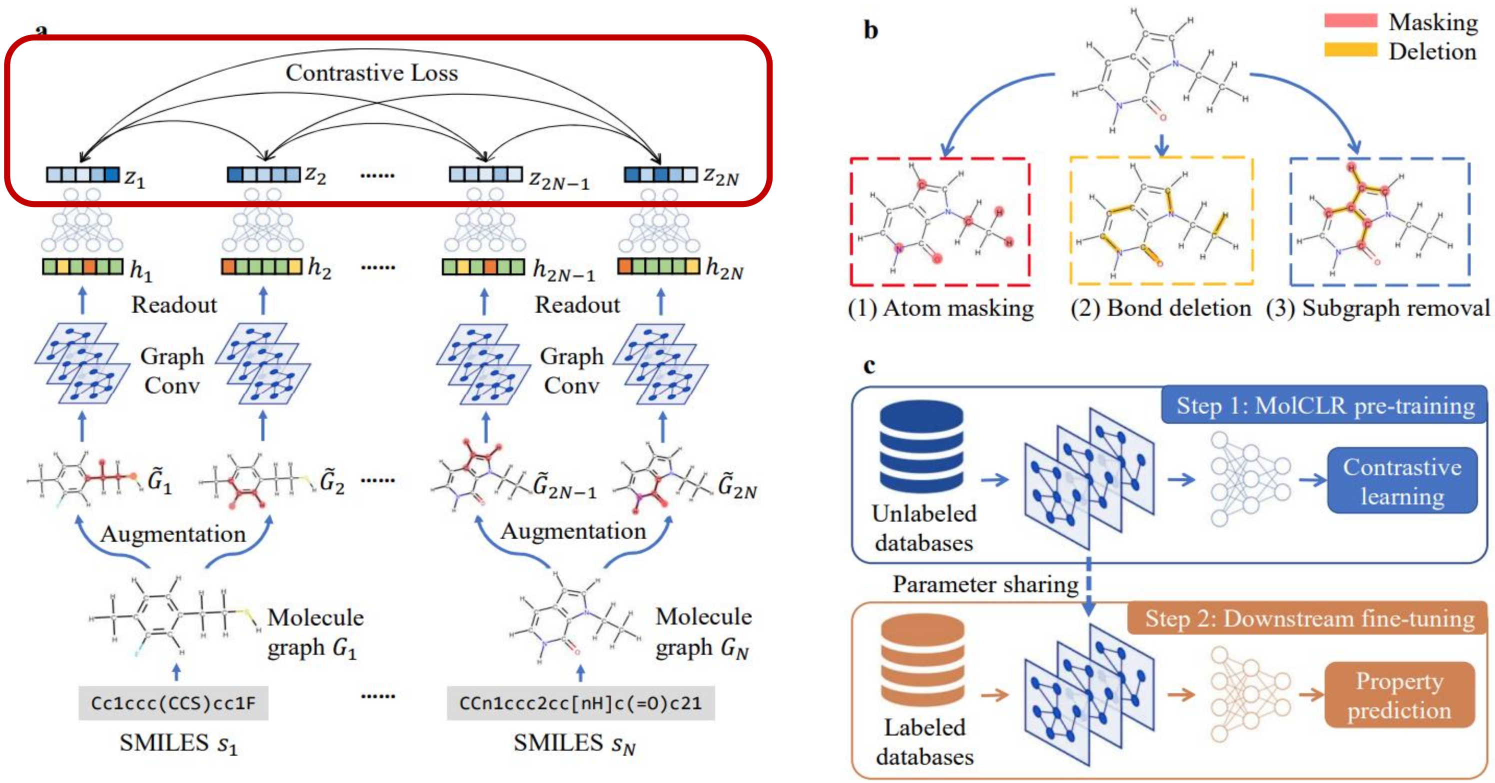
i^{+}: Positive

k: Negative Examples

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – InfoNCE Loss Function
  - ✓ MLP로 요약된 벡터들을 사용하여 대조 학습 수행

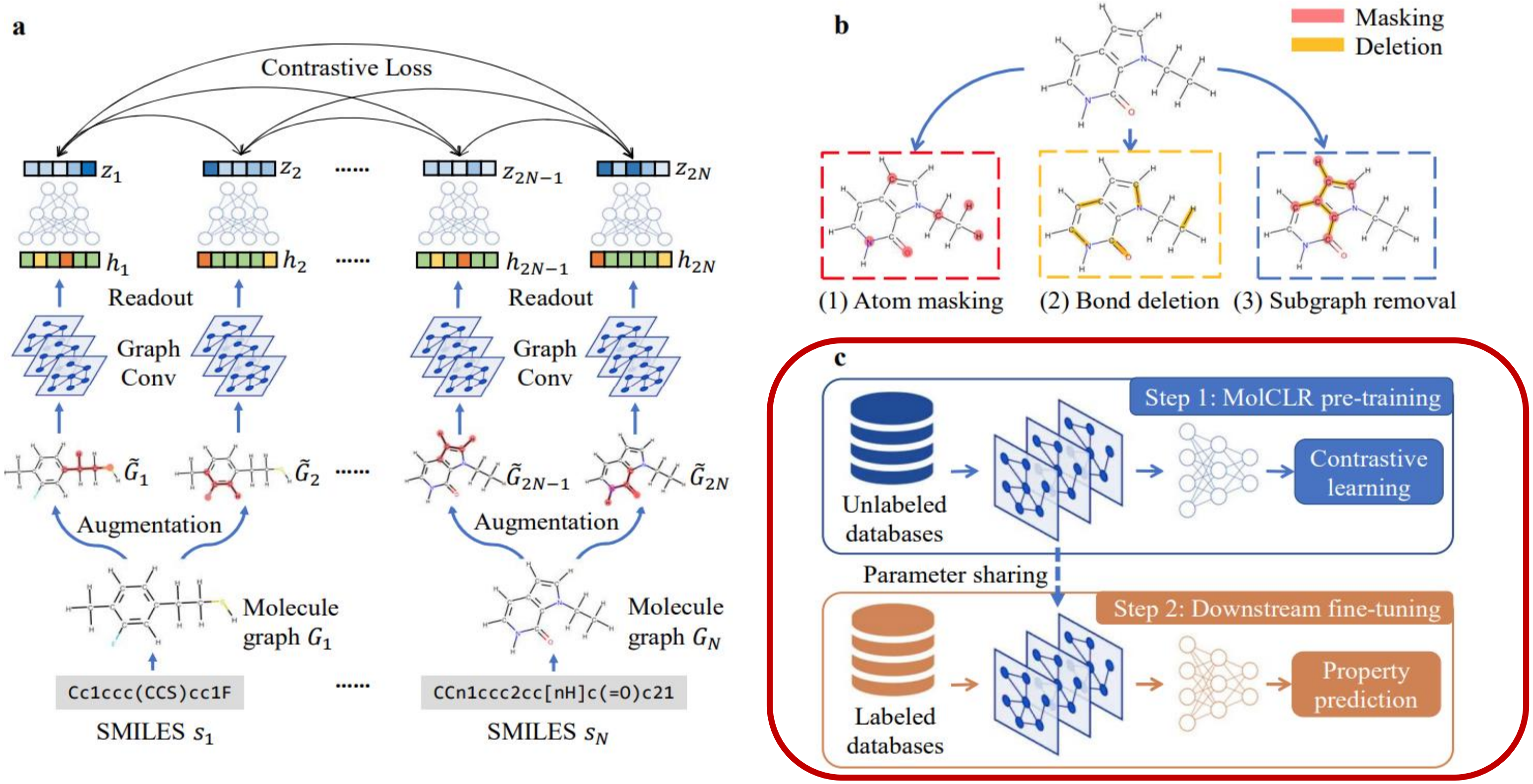


Applications of Self-Supervised Learning

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – 전체 학습 프로세스
  - ✓ Step 1: 제안한 MolCLR 방법으로 분자 그래프 표현 학습을 수행(Pre-Training)
  - ✓ Step 2: 학습된 그래프 인코더를 사용하여 분자 성질 예측 수행(Downstream Tasks)

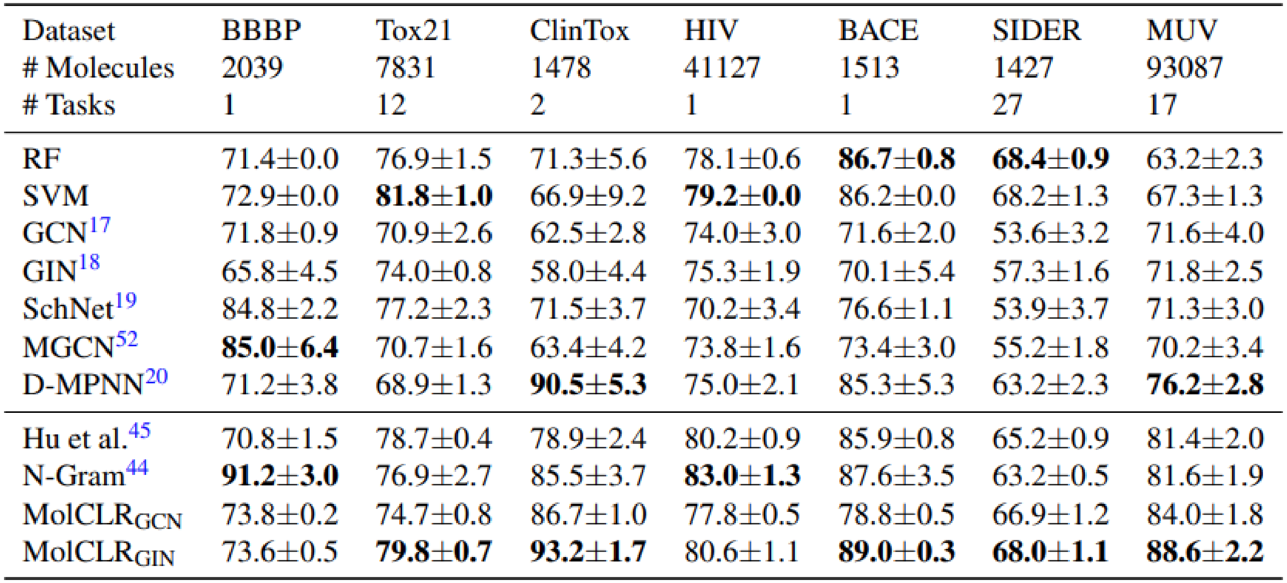


Applications of Self-Supervised Learning

Graph

❖ Experiments – Molecular Property Predictions

- Seven Classification Benchmarks (MoleculeNet) 사용
- Supervised Learning / Self-Supervised or Pre-Training Method에 대한 결과
- ROC-AUC (%)로 평가



	Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER	MUV
	# Molecules	2039	7831	1478	41127	1513	1427	93087
	# Tasks	1	12	2	1	1	27	17
Supervised Learning	RF	71.4±0.0	76.9±1.5	71.3±5.6	78.1±0.6	**86.7±0.8**	**68.4±0.9**	63.2±2.3
	SVM	72.9±0.0	**81.8±1.0**	66.9±9.2	**79.2±0.0**	86.2±0.0	68.2±1.3	67.3±1.3
	GCN[17]	71.8±0.9	70.9±2.6	62.5±2.8	74.0±3.0	71.6±2.0	53.6±3.2	71.6±4.0
	GIN[18]	65.8±4.5	74.0±0.8	58.0±4.4	75.3±1.9	70.1±5.4	57.3±1.6	71.8±2.5
	SchNet[19]	84.8±2.2	77.2±2.3	71.5±3.7	70.2±3.4	76.6±1.1	53.9±3.7	71.3±3.0
	MGCN[52]	**85.0±6.4**	70.7±1.6	63.4±4.2	73.8±1.6	73.4±3.0	55.2±1.8	70.2±3.4
	D-MPNN[20]	71.2±3.8	68.9±1.3	**90.5±5.3**	75.0±2.1	85.3±5.3	63.2±2.3	**76.2±2.8**
Self-Supervised or Pre-Training	Hu et al.[45]	70.8±1.5	78.7±0.4	78.9±2.4	80.2±0.9	85.9±0.8	65.2±0.9	81.4±2.0
	N-Gram[44]	**91.2±3.0**	76.9±2.7	85.5±3.7	**83.0±1.3**	87.6±3.5	63.2±0.5	81.6±1.9
	$\text{MolCLR}_{\text{GCN}}$	73.8±0.2	74.7±0.8	86.7±1.0	77.8±0.5	78.8±0.5	66.9±1.2	84.0±1.8
	$\text{MolCLR}_{\text{GIN}}$	73.6±0.5	**79.8±0.7**	**93.2±1.7**	80.6±1.1	**89.0±0.3**	**68.0±1.1**	**88.6±2.2**

Table 1. Test performance of different models on seven classification benchmarks. The first seven models are supervised learning methods and the last four are self-supervised/pre-training methods. Mean and standard deviation of test ROC-AUC (%) on each benchmark are reported.*

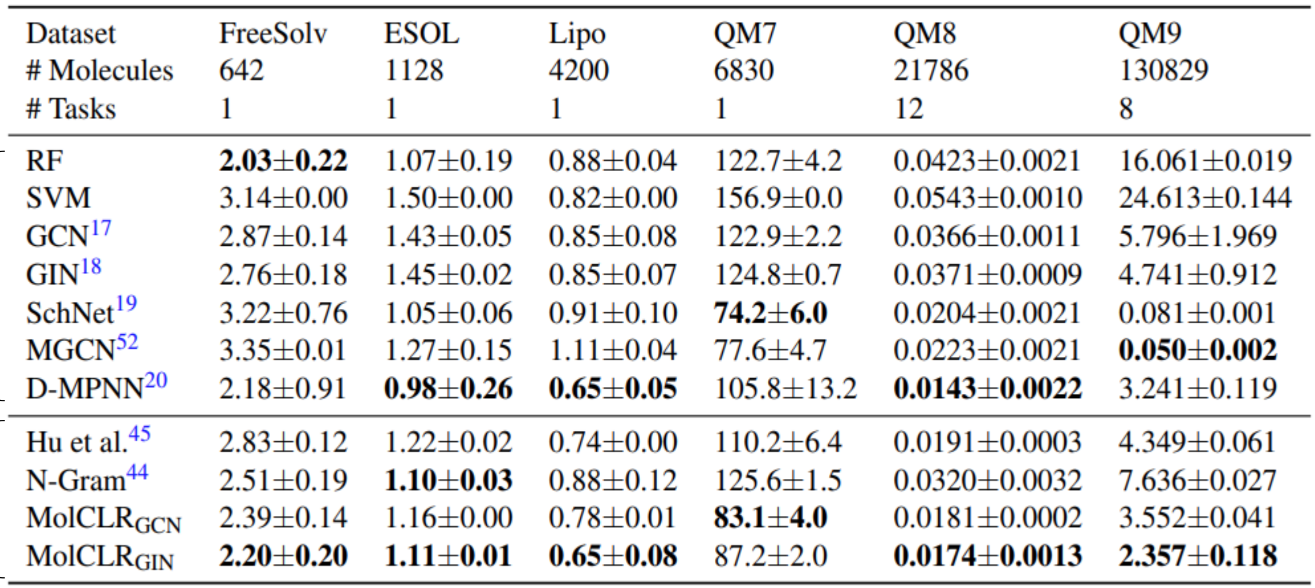
*Best performing supervised and self-supervised/pre-training methods for each benchmark are marked as bold.

Applications of Self-Supervised Learning

Graph

❖ Experiments – Molecular Property Predictions

- Six Regression Benchmarks (MoleculeNet) 사용
- Supervised Learning / Self-Supervised or Pre-Training Method에 대한 결과
- Root Mean Square Error (RMSE)로 평가



	Dataset	FreeSolv	ESOL	Lipo	QM7	QM8	QM9
	# Molecules	642	1128	4200	6830	21786	130829
	# Tasks	1	1	1	1	12	8
Supervised Learning	RF	**2.03±0.22**	1.07±0.19	0.88±0.04	122.7±4.2	0.0423±0.0021	16.061±0.019
	SVM	3.14±0.00	1.50±0.00	0.82±0.00	156.9±0.0	0.0543±0.0010	24.613±0.144
	GCN[17]	2.87±0.14	1.43±0.05	0.85±0.08	122.9±2.2	0.0366±0.0011	5.796±1.969
	GIN[18]	2.76±0.18	1.45±0.02	0.85±0.07	124.8±0.7	0.0371±0.0009	4.741±0.912
	SchNet[19]	3.22±0.76	1.05±0.06	0.91±0.10	**74.2±6.0**	0.0204±0.0021	0.081±0.001
	MGCN[52]	3.35±0.01	1.27±0.15	1.11±0.04	77.6±4.7	0.0223±0.0021	**0.050±0.002**
	D-MPNN[20]	2.18±0.91	**0.98±0.26**	**0.65±0.05**	105.8±13.2	**0.0143±0.0022**	3.241±0.119
Self-Supervised or Pre-Training	Hu et al.[45]	2.83±0.12	1.22±0.02	0.74±0.00	110.2±6.4	0.0191±0.0003	4.349±0.061
	N-Gram[44]	2.51±0.19	**1.10±0.03**	0.88±0.12	125.6±1.5	0.0320±0.0032	7.636±0.027
	$\text{MolCLR}_{\text{GCN}}$	2.39±0.14	1.16±0.00	0.78±0.01	**83.1±4.0**	0.0181±0.0002	3.552±0.041
	$\text{MolCLR}_{\text{GIN}}$	**2.20±0.20**	**1.11±0.01**	**0.65±0.08**	87.2±2.0	**0.0174±0.0013**	**2.357±0.118**

Table 2. Test performance of different models on six regression benchmarks. The first seven models are supervised learning methods and the last four are self-supervised/pre-training methods. Mean and standard deviation of test RMSE (for FreeSolv, ESOL, Lipo) or MAE (for QM7, QM8, QM9) are reported.*

*Best performing supervised and self-supervised/pre-training methods for each benchmark are marked as bold.

Conclusion

- ❖ 지도 학습 기반 분류 문제는 대용량의 데이터와 레이블 정보를 필요로 함
- ❖ 레이블이 없는 데이터를 효과적으로 표현할 수 있는 자기 지도 학습 등장
- ❖ 이미지 분류에 특화된 많은 자기 지도 학습 방법론이 제안됨
- ❖ 최근에는 이미지 뿐만 아니라 비디오, 의료, 오디오, 그래프, 강화학습 분야에서 자기 지도 학습 방법론들이 응용됨
- ❖ 연구하고 있는 분야, 데이터 형태, 도메인에 따라 학습 기법의 응용력을 높이는 계기가 되었으면 함

References

- He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).
- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.
- Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems, 33, 21271-21284.
- Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., & Langlotz, C. P. (2020). Contrastive learning of medical visual representations from paired images and text. arXiv preprint arXiv:2010.00747.
- Qian, R., Meng, T., Gong, B., Yang, M. H., Wang, H., Belongie, S., & Cui, Y. (2021). Spatiotemporal contrastive video representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6964-6974).
- Hansen, N., & Wang, X. (2021, May). Generalization in reinforcement learning by soft data augmentation. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13611-13617). IEEE.
- Wang, L., & Oord, A. V. D. (2021). Multi-format contrastive learning of audio representations. arXiv preprint arXiv:2103.06508.
- Wang, Y., Wang, J., Cao, Z., & Farimani, A. B. (2021). MolCLR: molecular contrastive learning of representations via graph neural networks. arXiv preprint arXiv:2102.10056.요~

- http://dmqa.korea.ac.kr/activity/seminar/341 (Multi-modal Learning)
- http://dmqa.korea.ac.kr/activity/seminar/319 (State Representation Learning for Reinforcement Learning)
- http://dmqa.korea.ac.kr/activity/seminar/310 (Dive into BYOL)
- http://dmqa.korea.ac.kr/activity/seminar/308 (Towards Contrastive Learning)
- http://dmqa.korea.ac.kr/activity/seminar/307 (The whys and hows of data augmentation)

Thank you